ARBITRARY-ACCURACY NEURAL APPROXIMATION WITH OPTIMAL NEURON COUNT AND NEAR-OPTIMAL BIT COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the minimum number of hidden neurons required for arbitrary-accuracy approximation of multivariate Hölder-continuous functions on $[0,1]^d$ and the associated encoding complexity. For $d \geq 2$, we construct a fixed, explicitly defined activation function for which a closed-form network with two hidden layers of widths d and 1 achieves arbitrary accuracy in the uniform norm. We prove that $d+1$ is the exact minimum total number of hidden neurons among standard feedforward networks with locally integrable activations and affine outputs. We further give a simpler construction using a single elementary activation that combines the floor and exponential functions. This construction requires three hidden layers of widths d, 1, and 2, only two neurons above the minimum. If a skip connection is allowed, widths d, 1, and 1 suffice. These constructions use explicit grid addressing and integer encoding of quantized function values. For a bounded α-Hölder class, they require $O(\varepsilon^{-d/\alpha}\log(1/\varepsilon))$ bits, matching the metric-entropy lower bound up to a logarithmic factor.

1 INTRODUCTION

Universal approximation theorems (Cybenko, 1989; Hornik et al., 1989; Hornik, 1991; Leshno et al., 1993) state that feedforward networks with suitable nonpolynomial activations are dense in $C([0,1]^d)$. In all classical statements, and in the quantitative theory that followed for ReLU networks (Yarotsky, 2017; 2018; Petersen & Voigtlaender, 2018; Shen et al., 2022), the number of neurons grows without an explicit bound as the prescribed accuracy ε tends to zero. A different regime was opened by Maiorov & Pinkus (1999): there is an analytic sigmoidal activation for which a two-hidden-layer network with widths $3d$ and $6d+3$ approximates every $f \in C([0,1]^d)$ to arbitrary accuracy, with the architecture fixed and only the parameters depending on f and ε. Such *fixed-size* universal approximators have since been constructed with progressively fewer neurons and progressively more explicit activations (Ismailov, 2014; Guliyev & Ismailov, 2018; Yarotsky, 2021; Shen et al., 2021; Zhang et al., 2022). In what follows, we list two questions that are left open by this line of work and will be answered here.

How few neurons suffice? The smallest known count is $3d+2$ hidden neurons, in two hidden layers of widths d and $2d+2$, obtained by Ismailov (2014), whose activation is a specially constructed C^∞ sigmoidal function whose existence is proved theoretically; Guliyev & Ismailov (2018) attain the same count and give an algorithmically computable construction. If different activations and skip connections are allowed, Yarotsky (2021) obtained a smaller construction with $d+2$ hidden neurons. No lower bound on the total number of hidden neurons was known beyond the trivial observation that the first layer must see the whole input. We construct and prove that **the exact answer is** $d+1$: a two-hidden-layer network of widths $(d,1)$ with a single explicit piecewise-constant activation and without any skip connection approximates every function in the Hölder class $\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$ (see (2)) to arbitrary L^∞ accuracy (Theorem 3.1), and no feedforward network with at most d hidden neurons can do so, whatever its depth and whatever locally integrable activation it uses (Theorem 3.4). The lower bound combines a rank argument for the first hidden layer with the fact, going back to (Lin & Pinkus, 1993; Maiorov & Pinkus, 1999; Pinkus, 1999), that sums of a fixed number of ridge functions are not dense; we give a direct proof valid for arbitrary locally integrable activations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of fixed-size universal approximators on $[0,1]^d$. Unless "two" is indicated, a single activation is used. *Optimal among all locally integrable activations for $d \geq 2$ (Theorem 3.4).

Reference	Neurons	Depth	Activation	Parameters
Maiorov & Pinkus (1999)	$9d+3$	2	analytic sigmoidal	existence
Ismailov (2014)	$3d+2$	2	C^∞ sigmoidal	existence
Guliyev & Ismailov (2018)	$3d+2$	2	computable C^∞ sigmoidal	existence
Yarotsky (2021)	$d+2$ (skip)	3	two: floor and analytic, elementary	existence
	$O(d^2)$	$O(1)$	two: sin and arcsin, elementary	existence
	$O(d^2)$	$O(1)$	C^1 sigmoidal, elementary	existence
Zhang et al. (2022)	$\leq 396d(2d+1)$	11	C^0 triangular wave–softsign, elementary	existence
Theorem 3.1	$d+1$ **(optimal*)**	2	explicit piecewise constant	closed-form
Theorem 4.1	$d+3$	3	floor–exponential, elementary	closed-form
Theorem 4.1	$d+2$ (skip)	3	floor–exponential, elementary	closed-form

What does a small network cost? Fixed-size constructions pay for accuracy with the magnitude and precision of their parameters, a phenomenon Zhang et al. (2022) call the *curse of memory* but do not quantify. We therefore define the *bit complexity* of a network as the total binary length of its (rational) parameters and show that our constructions have worst-case bit complexity $\Theta(\varepsilon^{-d/\alpha}\log(1/\varepsilon))$ over the Hölder class (Theorem 5.1). On the other hand, the classical Kolmogorov–Tikhomirov ε-*entropy* estimate (Kolmogorov & Tikhomirov, 1959) implies that, for any fixed architecture and activation, any family of rational-parameter networks approximating the entire class to accuracy ε must have worst-case bit complexity $\Omega(\varepsilon^{-d/\alpha})$ (Proposition 5.2). Thus, our constructions are information-theoretically near-optimal, up to a logarithmic factor. For fixed-size networks, bit complexity therefore provides a natural measure of representational cost.

Elementary activations The activation attaining $d+1$ neurons is explicit but intricate: it is a piecewise-constant function that lists, block by block, the digits of every integer in every base (see (10)). We therefore also consider a floor-exponential activation

$$\sigma(t) = \begin{cases} 2^t - 1, & t < 0, \\ \lfloor t \rfloor, & t \geq 0. \end{cases} \tag{1}$$

We call an activation *elementary* if it is given by a finite formula in finitely many standard functions; σ is elementary in this sense. Using σ as the single shared activation, we give an explicit three-hidden-layer network of widths $(d,1,2)$, hence $d+3$ hidden neurons, and a variant of widths $(d,1,1)$ with one skip connection, hence $d+2$ hidden neurons (Theorem 4.1). All parameters of both networks are rational and given in closed form, which allows exact-arithmetic verification (Section 6); to our knowledge, no earlier fixed-size construction has this property. Elementary activations have appeared in fixed-size constructions recently (Yarotsky, 2021; Zhang et al., 2022), but there the function-dependent weights are obtained from the density of an irrational winding on the torus and are not given in closed form; Yarotsky (2021) notes that computing the weight is practically infeasible even for small problems. Moreover, those earlier constructions need $O(d^2)$ neurons when a single elementary activation is used. Table 1 summarizes the results.

We summarize our **main contributions** as follows.

1. **Optimal neuron count.** We construct a network of widths $(d,1)$ with a single explicit activation ϱ in (10) and rational, closed-form parameters that approximates every $f \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$ to arbitrary L^∞ accuracy (Theorem 3.1), and we verify it in exact arithmetic (Section 6). We prove that $d+1$ is the exact minimum neuron count, at any depth and for any locally integrable activation (Theorem 3.4). We further show that, at least for $d=2$, no continuous activation attains this count in L^∞ (Proposition E.1), whereas under the weaker L^p criterion, $1 \leq p < \infty$, a continuous modification of ϱ that we construct attains it for every $d \geq 2$ (Theorem E.3).
2. **Elementary activation.** For the single elementary activation in (1), we give an explicit construction with $d+3$ hidden neurons and no skip connections, together with a $d+2$-neuron variant using one skip connection (Theorem 4.1). All parameters are rational and given in closed form, again permitting exact-arithmetic evaluation.
3. **Near-optimal storage complexity.** We develop the *bit complexity*, defined as the total binary length of the rational network parameters, to quantify the information stored in a fixed-size

network. Our constructions have worst-case bit complexity $\Theta(\varepsilon^{-d/\alpha}\log(1/\varepsilon))$ (Theorem 5.1), which is information-theoretically optimal up to a logarithmic factor.

The rest of the paper is organized as follows: Section 2 introduces the main concepts and notation. Section 3 establishes the optimal hidden-neuron count by proving both sufficiency and necessity. Section 4 presents the constructions based on the elementary floor–exponential activation. Section 5 analyzes the bit complexity of the proposed networks. Section 6 provides exact-arithmetic numerical verification, and Section 7 discusses limitations and open problems. Additional proofs, extensions, numerical results, and figures are provided in the appendices.

2 SETTING

To formulate precisely the minimum-neuron problem and quantify the associated encoding cost, we first introduce the function class, network setting, and bit complexity, followed by the grid-quantization scheme underlying all our constructions.

Hölder class. Throughout, $d \in \mathbb{N}_+$, $\alpha \in (0,1]$ and $\lambda, R > 0$. Consider

$$\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d) := \big\{ f \in C([0,1]^d) : \|f\|_{L^\infty([0,1]^d)} \le R,\ |f(\boldsymbol{x}) - f(\boldsymbol{y})| \le \lambda \|\boldsymbol{x}-\boldsymbol{y}\|^{\alpha}_{\infty},\ \forall \boldsymbol{x}, \boldsymbol{y} \big\}. \quad (2)$$

Since $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d) \subseteq \mathcal{H}^{\alpha}_{\lambda,\lceil R\rceil}([0,1]^d)$, we assume without loss of generality that $R \in \mathbb{N}_+$ whenever we need to make certain neural network parameters integers.

Networks. An *activation* is a function $\phi : \mathbb{R} \to \mathbb{R}$, applied componentwise to vectors. For the number of hidden layers $L \in \mathbb{N}_+$ and widths $\boldsymbol{n} = (n_1, \dots, n_L) \in \mathbb{N}_+^L$, a *feedforward network* with input dimension d, activation ϕ and architecture $\boldsymbol{n}$ is a function $\Phi : [0,1]^d \to \mathbb{R}$ of the form

$$h_0(\boldsymbol{x}) = \boldsymbol{x}, \quad h_k(\boldsymbol{x}) = \phi\big(W_k h_{k-1}(\boldsymbol{x}) + \boldsymbol{b}_k\big) \in \mathbb{R}^{n_k},\ k = 1, \dots, L, \quad \Phi(\boldsymbol{x}) = \boldsymbol{a}^\top h_L(\boldsymbol{x}) + c, \quad (3)$$

with $W_k \in \mathbb{R}^{n_k \times n_{k-1}}$ ($n_0 = d$), $\boldsymbol{b}_k \in \mathbb{R}^{n_k}$, $\boldsymbol{a} \in \mathbb{R}^{n_L}$, $c \in \mathbb{R}$. We denote the set of all such networks by $\mathcal{N}(\boldsymbol{n}; \phi)$. We also consider the larger class $\mathcal{N}^{\mathrm{skip}}(\boldsymbol{n}; \phi)$ in which the output layer is affine in all hidden layers, $\Phi(\boldsymbol{x}) = \sum_{k=1}^{L} \boldsymbol{a}_k^\top h_k(\boldsymbol{x}) + c$; the input still enters only through the first hidden layer. The number of *hidden neurons* is $|\boldsymbol{n}| := n_1 + \cdots + n_L$, the *depth* is L, and the *width* is $\max_k n_k$.

Definition 2.1 (Arbitrary-accuracy approximation and minimal neuron count)**.** The network class $\mathcal{N}(\boldsymbol{n}; \phi)$ *approximates* $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ *to arbitrary accuracy* if, for every $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ and every $\varepsilon > 0$, there exists $\Phi \in \mathcal{N}(\boldsymbol{n}; \phi)$ such that $\|f - \Phi\|_{L^\infty([0,1]^d)} \le \varepsilon$; the same definition applies to $\mathcal{N}^{\mathrm{skip}}(\boldsymbol{n}; \phi)$. The *minimal neuron count* of $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ is

$$\nu(\mathcal{H}) := \min\big\{ |\boldsymbol{n}| : \exists L \in \mathbb{N}_+,\ \boldsymbol{n} \in \mathbb{N}_+^L,\ \phi \in L^1_{\mathrm{loc}}(\mathbb{R}), \\ \forall f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d),\ \inf_{\Phi \in \mathcal{N}(\boldsymbol{n};\phi)} \|f - \Phi\|_{L^\infty([0,1]^d)} = 0 \big\}.$$

Only the architecture and the activation are fixed in Definition 2.1; the weights and biases may depend on f and ε. All activations used in our constructions are locally bounded and hence locally integrable.

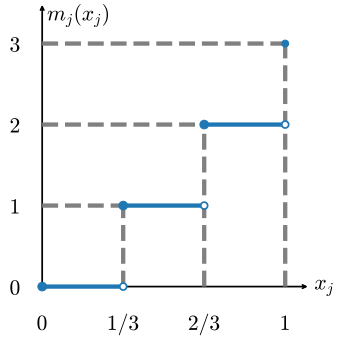
Bit complexity. For a nonnegative integer k let $\mathrm{bit}(k) := \lceil \log_2(k+1) \rceil$ be its binary length; and for a rational number $\theta = a/b$ in lowest terms ($a \in \mathbb{Z}$, $b \in \mathbb{N}_+$), let $\mathrm{bit}(\theta) := 1 + \mathrm{bit}(|a|) + \mathrm{bit}(b)$, the additional bit recording the sign. If all weights and biases of a network Φ are rational, define its *bit complexity* by

$$\mathrm{bit}(\Phi) := \sum_{\theta \in \mathcal{P}(\Phi)} \mathrm{bit}(\theta), \quad (4)$$

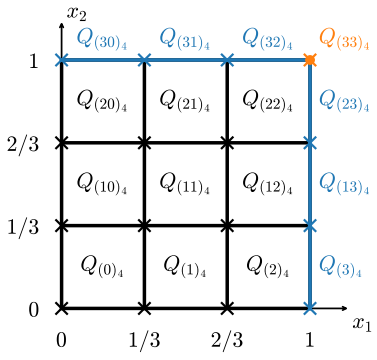
where $\mathcal{P}(\Phi)$ denotes the set of nonzero weights and biases of Φ.

Grid quantization (for all constructions). Fix $M \in \mathbb{N}_+$ and set $K := (M+1)^d$. For $\boldsymbol{x} \in [0,1]^d$ let $m_j(x_j) := \lfloor M x_j \rfloor \in \{0, \dots, M\}$ and define the *cell address*

$$r(\boldsymbol{x}) := \sum_{j=1}^{d} (M+1)^{j-1} m_j(x_j) \in \{0, \dots, K-1\}.$$



(a) $m_j(x_j) = \lfloor Mx_j \rfloor$ maps $[\frac{i}{M}, \frac{i+1}{M})$ to i and 1 to M.



(b) $r(\boldsymbol{x})$ for $d = 2$: each cell Q_ℓ is mapped to ℓ; crosses mark the representatives $\boldsymbol{x}^{(\ell)}$.

Figure 1: Position encoding with $M = 3$.

The address is the base-$(M+1)$ integer whose digits are the coordinates $m_j(x_j)$. For $\ell \in \{0, \dots, K-1\}$, the *cell* $Q_\ell := \{\boldsymbol{x} \in [0,1]^d : r(\boldsymbol{x}) = \ell\}$ is a half-open box (degenerate on $x_j = 1$) and can be represented by the grid point $\boldsymbol{x}^{(\ell)} := \big(m_1^{(\ell)}/M, \dots, m_d^{(\ell)}/M\big)$ whose coordinates $m_j^{(\ell)} = \lfloor \ell/(M+1)^{j-1} \rfloor \bmod (M+1)$ are the digits of ℓ. See Figure 1 and Table 2 in Appendix F for an example of $M = 3$. Then $\|\boldsymbol{x} - \boldsymbol{x}^{(\ell)}\|_\infty \le 1/M$ for all $\boldsymbol{x} \in Q_\ell$. Given $\varepsilon > 0$, we always choose

$$\delta := 2^{-\lceil \log_2(2/\varepsilon) \rceil} \in \big(\tfrac{\varepsilon}{4}, \tfrac{\varepsilon}{2}\big], \quad M := \Big\lceil \big(\tfrac{2\lambda}{\varepsilon}\big)^{1/\alpha} \Big\rceil, \quad B := \max\big\{2,\, 2^{\lceil \log_2(\lfloor 2R/\delta \rfloor + 1) \rceil}\big\}, \tag{5}$$

so that $\lambda M^{-\alpha} \le \varepsilon/2$, δ is dyadic and B is a power of two. For $f \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$, we define the quantization indices q_ℓ and the corresponding quantized approximations $\widehat{f}_\ell$ of $f(\boldsymbol{x}^{(\ell)})$ by

$$q_\ell := \Big\lfloor \frac{f(\boldsymbol{x}^{(\ell)}) + R}{\delta} \Big\rfloor \in \{0, \dots, B-1\}, \quad \widehat{f}_\ell := -R + \delta q_\ell, \quad 0 \le f(\boldsymbol{x}^{(\ell)}) - \widehat{f}_\ell < \delta, \tag{6}$$

which satisfy, for every $\boldsymbol{x} \in Q_\ell$,

$$|f(\boldsymbol{x}) - \widehat{f}_\ell| \le |f(\boldsymbol{x}) - f(\boldsymbol{x}^{(\ell)})| + |f(\boldsymbol{x}^{(\ell)}) - \widehat{f}_\ell| < \lambda M^{-\alpha} + \delta \le \varepsilon. \tag{7}$$

The construction below is a network that outputs $\widehat{f}_{r(\boldsymbol{x})}$, or equivalently $\widehat{f}_\ell$ when $\boldsymbol{x} \in Q_\ell$. Moreover,

$$M = \Theta(\varepsilon^{-1/\alpha}), \quad K = \Theta(\varepsilon^{-d/\alpha}), \quad B = \Theta(\varepsilon^{-1}) \quad \text{as } \varepsilon \to 0. \tag{8}$$

3 Optimal neuron count: $d + 1$ hidden neurons

3.1 Sufficiency: an explicit $(d, 1)$ network

Our goal is to realize the piecewise-constant approximation $\widehat{f}_{r(\boldsymbol{x})} = -R + \delta q_{r(\boldsymbol{x})}$. Our strategy is first to determine the cell containing $\boldsymbol{x}$ through its address $r(\boldsymbol{x})$, and then to pack $q_0, \dots, q_{K-1}$ into a single integer as digits in a suitable base, so that $q_{r(\boldsymbol{x})}$ can be recovered by extracting the digit indexed by $r(\boldsymbol{x})$. This motivates an activation ϱ whose negative branch provides the floor operation needed for addressing and whose nonnegative branch performs digit extraction.

The digit-extraction activation. For $n \in \mathbb{N}_+$, let $S_1 := 0$ and $S_n := \sum_{k=1}^{n-1} k(k+1)^k$, so that $S_{n+1} - S_n = n(n+1)^n$ and $S_n \to \infty$. The half-line $[0, \infty)$ is the disjoint union of the blocks $[S_n, S_{n+1})$, $n \in \mathbb{N}_+$, and the n-th block is the disjoint union of the unit intervals

$$I_{n,j,r} := [S_n + nj + r, S_n + nj + r + 1), \quad j \in \{0, \dots, (n+1)^n - 1\}, \quad r \in \{0, \dots, n-1\}. \tag{9}$$

Write $j = \sum_{s=0}^{n-1} b_s(j)(n+1)^{n-1-s}$ with digits $b_s(j) \in \{0, \dots, n\}$. Define

$$\varrho(t) := \begin{cases} \lfloor t \rfloor, & t < 0, \\ b_r(j), & t \in I_{n,j,r}. \end{cases} \tag{10}$$

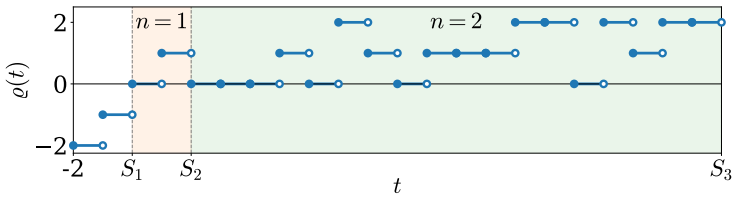


Figure 2: ϱ in $[-2, S_3)$, or equivalently, $[-2, 0) \bigcup \big(\bigcup_{n=1}^{2} \bigcup_{j=0}^{(n+1)^n-1} \bigcup_{r=0}^{n-1} I_{n,j,r} \big)$.

Equivalently, on $I_{n,j,r}$ the value $b_r(j)$ is $\lfloor j/(n+1)^{n-1-r} \rfloor - (n+1)\lfloor j/(n+1)^{n-r} \rfloor$. As shown in Figure 2,the function ϱ is piecewise constant on countably many intervals, satisfies $0 \le \varrho(t) \le n$ on $[S_n, S_{n+1})$, and is therefore locally bounded and locally integrable. Its defining property is the *digit-extraction identity*

$$\varrho(S_n + nj + r) = b_r(j), \quad j \in \{0, \dots, (n+1)^n - 1\}, \quad r \in \{0, \dots, n-1\}. \tag{11}$$

The network. Let $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$, $\varepsilon > 0$, and let M, δ, B, q_ℓ be as in (5)–(6). Set

$$n := \max\{K, B-1\}, \quad J_f := \sum_{\ell=0}^{K-1} q_\ell \, (n+1)^{n-1-\ell}, \quad N_f := S_n + nJ_f. \tag{12}$$

Since $q_\ell \le B-1 \le n$, the base-$(n+1)$ digits of J_f are $b_\ell(J_f) = q_\ell$ for $\ell < K$ and 0 for $K \le \ell < n$; in particular $0 \le J_f < (n+1)^n$, and (11) gives $\varrho(N_f + \ell) = q_\ell$ for $0 \le \ell < K$.

As illustrated in Figure 3, the outputs of the hidden layers and the affine output layer are defined as:

$$u_j(\boldsymbol{x}) := \varrho\big(Mx_j - (M+1)\big) = \lfloor Mx_j \rfloor - (M+1), \quad j = 1, \dots, d, \quad \text{(1st hidden layer)} \tag{13}$$

$$z(\boldsymbol{x}) := \varrho\Big(N_f + \sum_{j=1}^{d} (M+1)^{j-1} \big(u_j(\boldsymbol{x}) + M + 1\big)\Big) \quad \text{(2nd hidden layer)}$$

$$= \varrho\big(N_f + r(\boldsymbol{x})\big) = q_{r(\boldsymbol{x})} \tag{14}$$

$$\Phi^{\varrho}_{f,\varepsilon}(\boldsymbol{x}) := -R + \delta z(\boldsymbol{x}) = \widehat{f}_{r(\boldsymbol{x})}. \quad \text{(output layer)} \tag{15}$$

In (13) the argument of ϱ is at most -1, so the negative branch (the floor) applies; in (14) the argument is $N_f + r(\boldsymbol{x}) \in [S_n, S_{n+1})$ and the digit-extraction branch applies. Since the pointwise estimate (7) holds on every Q_ℓ and $\bigcup_{\ell=0}^{K-1} Q_\ell = [0,1]^d$, taking the supremum over $\boldsymbol{x} \in [0,1]^d$ yields:

Theorem 3.1 (Two hidden layers: $d+1$ neurons)**.** *Let $d \in \mathbb{N}_+$, $\alpha \in (0,1]$, $\lambda, R > 0$. For every $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ and every $\varepsilon > 0$ the network $\Phi^{\varrho}_{f,\varepsilon} \in \mathcal{N}\big((d,1);\varrho\big)$ defined by* (13)–(15) *satisfies*

$$\|f - \Phi^{\varrho}_{f,\varepsilon}\|_{L^\infty([0,1]^d)} \le \varepsilon.$$

In particular $\nu(\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)) \le d+1$.

All parameters (see details in Appendix A) are integers except the dyadic output weight δ and the output bias $-R$; the latter is also an integer if, without loss of generality, we take $R \in \mathbb{N}_+$. The only function-dependent quantity among the parameters is N_f, which enters the second-layer bias and encodes the quantized values of f required for accuracy ε. For fixed d, the architecture is independent of f and ε, while the activation ϱ is independent of all problem parameters.

3.2 Necessity: no network with d hidden neurons is universal

In this section, we prove that at least $d+1$ hidden neurons are necessary to approximate $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ to arbitrary accuracy in the uniform norm when $d \ge 2$. The proof relies on two activation-independent results: Proposition 3.2 shows that the first hidden layer must contain at least d neurons, and Proposition 3.3 shows that a single-hidden-layer network of fixed width is insufficient.

Proposition 3.2 (First-layer width obstruction)**.** *Let $d \ge 2$, $\alpha \in (0,1]$, $\lambda, R > 0$, and let $\phi : \mathbb{R} \to \mathbb{R}$ be arbitrary. For every architecture $\boldsymbol{n} = (n_1, \dots, n_L)$ with $n_1 < d$, there exist a function $f_0 \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ and a constant $\varepsilon_0 > 0$, depending only on α, λ, and R, such that*

$$\inf_{\Phi \in \mathcal{N}^{\mathrm{skip}}(\boldsymbol{n};\phi)} \|f_0 - \Phi\|_{L^\infty([0,1]^d)} \ge \varepsilon_0.$$

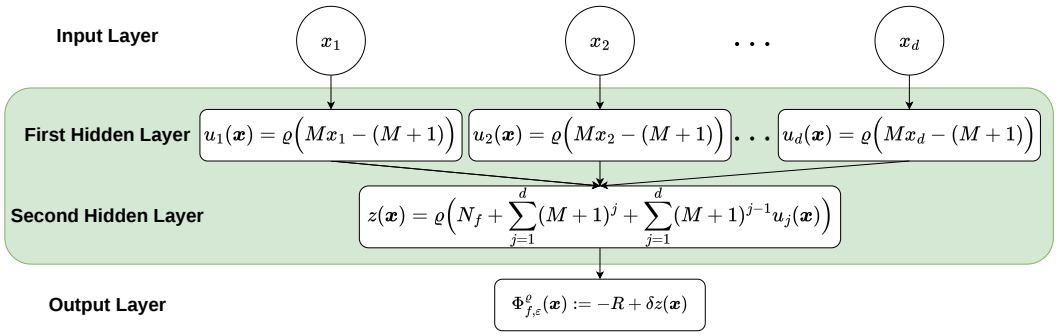


Figure 3: Architecture of the network $\Phi^{\varrho}_{f,\varepsilon}$ in Theorem 3.1. The d neurons in the first hidden layer quantize the input coordinates, whose weighted combination forms the cell address $r(\boldsymbol{x})$; the neuron in the second hidden layer decodes the corresponding quantized value from the stored integer N_f. The network has $d+1$ hidden neurons in total.

See the proof in Appendix B.1.

Proposition 3.3 (Fixed-width single-hidden-layer obstruction)**.** *Let $d \geq 2$, $\alpha \in (0,1]$, $\lambda, R > 0$, and $n_1 \in \mathbb{N}_+$, and let $\phi \in L^1_{\mathrm{loc}}(\mathbb{R})$. There exists a function $f_h \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ and a constant $\varepsilon_1 > 0$, depending only on d, λ, R, and n_1, such that*

$$\inf_{\Phi \in \mathcal{N}((n_1);\phi)} \|f_h - \Phi\|_{L^\infty([0,1]^d)} \geq \varepsilon_1.$$

In particular, no single-hidden-layer network of fixed width can approximate $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ to arbitrary accuracy.

See the proof in Appendix B.2. In fact, that fixed-width single-hidden-layer networks are never dense in $C([0,1]^d)$, $d \geq 2$, is classical for continuous activations: it follows from the non-density of sums of m ridge functions (Lin & Pinkus, 1993; Maiorov & Pinkus, 1999; Pinkus, 1999). We still include a direct proof because it is valid for arbitrary locally integrable ϕ.

Combining the results of the above two propositions, we obtain the following theorem:

Theorem 3.4 (Lower bound of neurons)**.** *Let $d \geq 2$, $\alpha \in (0,1]$, and $\lambda, R > 0$. Let $L \in \mathbb{N}_+$, $\boldsymbol{n} \in \mathbb{N}_+^L$, and $\phi \in L^1_{\mathrm{loc}}(\mathbb{R})$. If $\mathcal{N}^{\mathrm{skip}}(\boldsymbol{n};\phi)$ approximates $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ to arbitrary accuracy in the sense of Definition 2.1, then $|\boldsymbol{n}| \geq d+1$. Consequently, together with Theorem 3.1,*

$$\nu\big(\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)\big) = d+1 \quad \textit{for every } d \geq 2,\ \alpha \in (0,1],\ \lambda, R > 0.$$

See the proof in Appendix B.3. The hypotheses $\phi \in L^1_{\mathrm{loc}}(\mathbb{R})$ and $d \geq 2$ are essential; see the two remarks below.

Remark 3.5 (The case $d = 1$). The above results (Propositions 3.2–3.3 and Theorem 3.4) need $d \geq 2$. For $d = 1$, Guliyev & Ismailov (2016) construct an activation for which a single hidden neuron approximates every $f \in C([0,1])$ to arbitrary accuracy, so $\nu = 1$ when $d = 1$.

Remark 3.6 (Continuous activations and L^p approximation). The discontinuity of the activation cannot be dispensed with for uniform approximation with the optimal neuron count $d+1$ established in Theorems 3.1 and 3.4, at least when $d = 2$: Proposition E.1 shows that no network with three hidden neurons, a continuous activation, and no skip connections can achieve arbitrary accuracy in $L^\infty([0,1]^2)$. If the error criterion is weakened to $L^p([0,1]^d)$, $1 \leq p < \infty$, however, a $(d,1)$ architecture with a continuous activation can achieve arbitrary accuracy (see Theorem E.3).

4 Elementary activation: $d+3$ neurons (or $d+2$ with a skip)

The activation ϱ in Theorem 3.1 performs digit extraction in a single step by encoding every possible digit table in its graph. We now replace this lookup mechanism with elementary radix arithmetic. Store the quantized values as the base-B integer $A_f = \sum_\ell q_\ell B^\ell$. Its ℓ-th digit can be recovered by $q_\ell = \lfloor A_f/B^\ell \rfloor - B\lfloor A_f/B^{\ell+1} \rfloor$. Consequently, retrieving the digit indexed by $r(\boldsymbol{x})$ requires

computing $B^{-r(\boldsymbol{x})}$ and applying the floor function twice. This motivates an elementary activation that combines exponential and floor operations as in (1). Let $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$, $\varepsilon > 0$, M, K, δ, B, q_ℓ as in (5)–(6), and

$$A_f := \sum_{\ell=0}^{K-1} q_\ell B^\ell \in \{0, \dots, B^K - 1\}, \quad T_0(\boldsymbol{x}) := \Big\lfloor \frac{A_f}{B^{r(\boldsymbol{x})}} \Big\rfloor, \quad T_1(\boldsymbol{x}) := \Big\lfloor \frac{A_f}{B^{r(\boldsymbol{x})+1}} \Big\rfloor, \tag{16}$$

so that $T_0(\boldsymbol{x}) - BT_1(\boldsymbol{x}) = q_{r(\boldsymbol{x})}$ (all base-B digits at positions greater than $r(\boldsymbol{x})$ cancel). Therefore, we define the outputs of the hidden layers and the affine output layer of the elementary activation neural network with $\boldsymbol{c} := (1, M+1, \dots, (M+1)^{d-1})^\top$,

$$\boldsymbol{u}(\boldsymbol{x}) := \sigma(M\boldsymbol{x}) = \big(\lfloor Mx_1 \rfloor, \dots, \lfloor Mx_d \rfloor\big)^\top, \qquad \text{(1st hidden layer)} \tag{17}$$

$$v(\boldsymbol{x}) := \sigma\big(-1 - \log_2(B)\, \boldsymbol{c}^\top \boldsymbol{u}(\boldsymbol{x})\big) = \tfrac{1}{2} B^{-r(\boldsymbol{x})} - 1, \qquad \text{(2nd hidden layer)} \tag{18}$$

$$\boldsymbol{w}(\boldsymbol{x}) := \sigma\Big(\begin{bmatrix} 2A_f \\ 2A_f/B \end{bmatrix} \big(v(\boldsymbol{x}) + 1\big)\Big) = \begin{bmatrix} T_0(\boldsymbol{x}) \\ T_1(\boldsymbol{x}) \end{bmatrix}, \qquad \text{(3rd hidden layer)} \tag{19}$$

$$\Phi^{\sigma}_{f,\varepsilon}(\boldsymbol{x}) := -R + \delta\,[1, -B]\,\boldsymbol{w}(\boldsymbol{x}) = -R + \delta\, q_{r(\boldsymbol{x})} = \widehat{f}_{r(\boldsymbol{x})}. \qquad \text{(output layer)} \tag{20}$$

Here $Mx_j \geq 0$ in (17) selects the floor branch of $\sigma(t)$; the pre-activation part in (18) is ≤ -1 and selects the exponential branch, and $\log_2 B \in \mathbb{N}$ because B is a power of two; the pre-activations in (19) are nonnegative and select the floor branch. The network has widths $(d, 1, 2)$ (see Figure 6).

Saving one neuron with a skip connection. The first neuron at the third hidden layer (19) can be dispensed with if the output layer may read the second hidden layer directly, because $A_f/B^{r(\boldsymbol{x})}$ itself differs from $\lfloor A_f/B^{r(\boldsymbol{x})} \rfloor$ by the fractional part $\eta_{r(\boldsymbol{x})} := A_f B^{-r(\boldsymbol{x})} - \lfloor A_f B^{-r(\boldsymbol{x})} \rfloor \in [0,1)$, whose contribution $\delta\eta_{r(\boldsymbol{x})}$ to the output is below the quantization step. Thus, keeping (17)–(18) and setting

$$w^{\mathrm{skip}}(\boldsymbol{x}) := \sigma\big(\tfrac{2A_f}{B}(v(\boldsymbol{x}) + 1)\big) = T_1(\boldsymbol{x}), \quad \Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}(\boldsymbol{x}) := -R + 2\delta A_f\big(v(\boldsymbol{x}) + 1\big) - \delta B\, w^{\mathrm{skip}}(\boldsymbol{x}), \tag{21}$$

one obtains $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}(\boldsymbol{x}) = -R + \delta\big(A_f B^{-r(\boldsymbol{x})} - BT_1(\boldsymbol{x})\big) = \widehat{f}_{r(\boldsymbol{x})} + \delta\eta_{r(\boldsymbol{x})}$, a network in $\mathcal{N}^{\mathrm{skip}}((d,1,1);\sigma)$ with $d+2$ hidden neurons (see Figure 7).

These two networks also provide uniform approximation of $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$:

Theorem 4.1 (Elementary activation: $d+3$ neurons, or $d+2$ with a skip connection)**.** *Let $d \in \mathbb{N}_+$, $\alpha \in (0,1]$, λ, $R > 0$. For every $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ and every $\varepsilon > 0$, the networks $\Phi^{\sigma}_{f,\varepsilon} \in \mathcal{N}((d,1,2);\sigma)$ and $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon} \in \mathcal{N}^{\mathrm{skip}}((d,1,1);\sigma)$, defined respectively by* (17)–(20) *and* (17), (18), (21) *satisfy*

$$\big\|f - \Phi^{\sigma}_{f,\varepsilon}\big\|_{L^\infty([0,1]^d)} \leq \varepsilon \quad \textit{and} \quad \big\|f - \Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}\big\|_{L^\infty([0,1]^d)} \leq \varepsilon.$$

See the proof in Appendix C. Compared with the first construction of Yarotsky (2021), which also uses $d+2$ neurons and a skip connection, $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$ requires only a single shared activation, and all of its parameters are rational and given in closed form (see Table 3).

5 Bit Complexity of Fixed-Size Approximators

In each of the three constructions, the target function is encoded by a single integer, namely, N_f in (12) or A_f in (16). Although the number of neurons is fixed, the size of this integer increases as the prescribed accuracy improves. A natural measure of storage and computational cost is therefore the bit complexity defined in (4). This complexity is readily computed because all our network parameters are rational: δ is dyadic, B is a power of two, and the remaining quantities are integers for $R \in \mathbb{N}_+$. The preceding discussion leads to the following theorem:

Theorem 5.1 (Bit complexity of the constructions)**.** *Let $d \geq 2$, $\alpha \in (0,1]$, $\lambda > 0$, and $R \in \mathbb{N}_+$. For the networks $\Phi^{\varrho}_{f,\varepsilon}$, $\Phi^{\sigma}_{f,\varepsilon}$, and $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$ in Theorems 3.1 and 4.1, as $\varepsilon \to 0$,* $\sup_{f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(\Phi^{\varrho}_{f,\varepsilon}\big) = \Theta\big(\varepsilon^{-d/\alpha} \log \tfrac{1}{\varepsilon}\big)$, $\sup_{f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(\Phi^{\sigma}_{f,\varepsilon}\big) = \Theta\big(\varepsilon^{-d/\alpha} \log \tfrac{1}{\varepsilon}\big)$, $\sup_{f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}\big) = \Theta\big(\varepsilon^{-d/\alpha} \log \tfrac{1}{\varepsilon}\big)$.

The proof is given in Appendix D.

We next compare this complexity with the information-theoretic lower bound. The classical Kolmogorov–Tikhomirov entropy estimate (Kolmogorov & Tikhomirov, 1959, Theorem XIV, equation (68)) states that the *ε-entropy* of $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ in the uniform norm, defined as the base-2 logarithm of the minimum cardinality of an ε-cover of $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$, is $\Theta(\varepsilon^{-d/\alpha})$. Equivalently, the minimum cardinality of such a cover is $2^{\Theta(\varepsilon^{-d/\alpha})}$. For a fixed architecture, any collection of N distinct rational-parameter networks has worst-case bit complexity $\Omega(\log_2 N)$. Consequently, any family of networks approximating the entire class to accuracy ε must have worst-case bit complexity $\Omega(\varepsilon^{-d/\alpha})$. This yields the following result:

Proposition 5.2 (Bit complexity lower bound)**.** *Let $d \in \mathbb{N}_+$, $\alpha \in (0,1]$, $\lambda, R > 0$. Fix any $L \in \mathbb{N}_+$, $\boldsymbol{n} \in \mathbb{N}_+^L$, and activation $\phi : \mathbb{R} \to \mathbb{R}$. For each $\varepsilon > 0$ and $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$, let $\Phi^{\phi}_{f,\varepsilon} \in \mathcal{N}^{\mathrm{skip}}(\boldsymbol{n};\phi)$ be any network with rational parameters satisfying $\|f - \Phi^{\phi}_{f,\varepsilon}\|_{L^\infty([0,1]^d)} \le \varepsilon$. Then $\sup_{f\in\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)} \mathrm{bit}(\Phi^{\phi}_{f,\varepsilon}) = \Omega(\varepsilon^{-d/\alpha})$ as $\varepsilon \to 0$.*

Combining Proposition 5.2 with Theorem 5.1, our constructions are optimal up to a logarithmic factor. In particular, the polynomial storage requirement $\varepsilon^{-d/\alpha}$ is unavoidable for any finite-bit implementation of a fixed-size universal approximator. The "curse of memory" identified by Zhang et al. (2022) is therefore not a defect of a particular construction but an information-theoretic necessity.

The inadequacy of neuron count alone is also reflected in VC dimension. Indeed, consider $\mathcal{N}((d,1);\varrho)$ and any finite set of distinct points $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N \in [0,1]^d$ with arbitrary labels $y_1,\dots,y_N \in \{-1,1\}$. There exists a bounded Lipschitz function f such that $f(\boldsymbol{x}_i) = y_i$ for every i. By Theorem 3.1, there is a network $\Phi \in \mathcal{N}((d,1);\varrho)$ satisfying $\|f - \Phi\|_{L^\infty([0,1]^d)} \le 1/2$, and hence $y_i\Phi(\boldsymbol{x}_i) > 0$, $i = 1,\dots,N$. Thus, the classifier $\operatorname{sign}\Phi$ realizes the prescribed labels. Since the points and labels were arbitrary, $\mathcal{N}((d,1);\varrho)$ has infinite VC dimension. The same argument applies to the network classes in Theorem 4.1. Hence, for these superexpressive activations, the number of neurons or parameters alone does not control statistical complexity.

6 Numerical verification in exact arithmetic

The explicit construction of $\Phi^{\varrho}_{f,\varepsilon}$ in Theorem 3.1 permits exact-arithmetic evaluation. We carry this out for $d = 2, 3$ to provide a computational check of the construction.

We consider four functions $f_1,\dots,f_4$ on $[0,1]^2$ and two functions f_5, f_6 on $[0,1]^3$, with certified Hölder parameters (α,λ,R) in the ℓ^∞ norm:

$$
\begin{aligned}
f_1(\boldsymbol{x}) &= \tfrac{1}{2}\sin(\pi x_1)\cos(\pi x_2), && (\alpha,\lambda,R) = (1,\pi,1);\\
f_2(\boldsymbol{x}) &= |x_1 - \tfrac{1}{2}|^{1/2} + \tfrac{1}{2}|x_2 - x_1|^{1/2}, && (\alpha,\lambda,R) = (\tfrac{1}{2},\tfrac{7}{4},2);\\
f_3(\boldsymbol{x}) &= \exp\big[\sin(\pi p)\cos(\pi p)\big]\log\big[1 + \tfrac{p^2}{2+p}\big] + \tfrac{\sin[3\pi(x_1+x_2)]}{1+p^2}, && (\alpha,\lambda,R) = (1,19,1);\\
f_4(\boldsymbol{x}) &= e^{-3\,\mathrm{dist}(\boldsymbol{x},\Gamma)}\cos\big(4\pi\,\mathrm{dist}(\boldsymbol{x},\Gamma)\big), && (\alpha,\lambda,R) = (1,18.3,1);\\
f_5(\boldsymbol{x}) &= \tfrac{1}{2}\sin(\pi x_1 x_2)\cos(\pi x_3) + \tfrac{1}{4}\big|x_2 - x_1 x_3\big|, && (\alpha,\lambda,R) = (1,\tfrac{39}{10},1);\\
f_6(\boldsymbol{x}) &= \tfrac{1}{2}\cos(\pi\,\|\boldsymbol{x} - \boldsymbol{c}\|_2), \quad \boldsymbol{c} = (0.3,0.6,0.2), && (\alpha,\lambda,R) = (1,\tfrac{11}{4},1).
\end{aligned}
$$

For f_3, $p := x_1x_2$. For f_4, Γ denotes the boundary of the level-3 Koch snowflake obtained by applying three iterations of the classical outward Koch construction (Koch, 1904) to an upward-pointing equilateral triangle centered at $(\frac{1}{2},\frac{1}{2})$ with circumradius 0.36, and define $\mathrm{dist}(\boldsymbol{x},\Gamma) := \min_{\boldsymbol{y}\in\Gamma}\|\boldsymbol{x} - \boldsymbol{y}\|_2$.

We perform the verification over a sequence of dyadic accuracies starting from $\varepsilon = 2^{-1}$, with the finest tested accuracy chosen so that $\mathrm{bit}(N_f)$ is on the order of 10^7 bits. Table 3 in Appendix F reports the parameters $\delta, M, K = (M+1)^d, B$ from (5), the bit length $\mathrm{bit}(N_f)$ from (12), and the maximal error $\max|f_i - \Phi^{\varrho}_{f_i,\varepsilon}|$ over 3,004 exactly represented test points for $d = 2$ and 3,008 for $d = 3$, comprising 2,000 random dyadic points, 1,000 boundary points, and the 2^d corners. Figure 4 shows the target f_4, the network outputs $\Phi^{\varrho}_{f_4,\varepsilon}$, and the pointwise errors on a 161^2 display grid at the three common accuracy levels $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}$, together with the maximal errors over the full

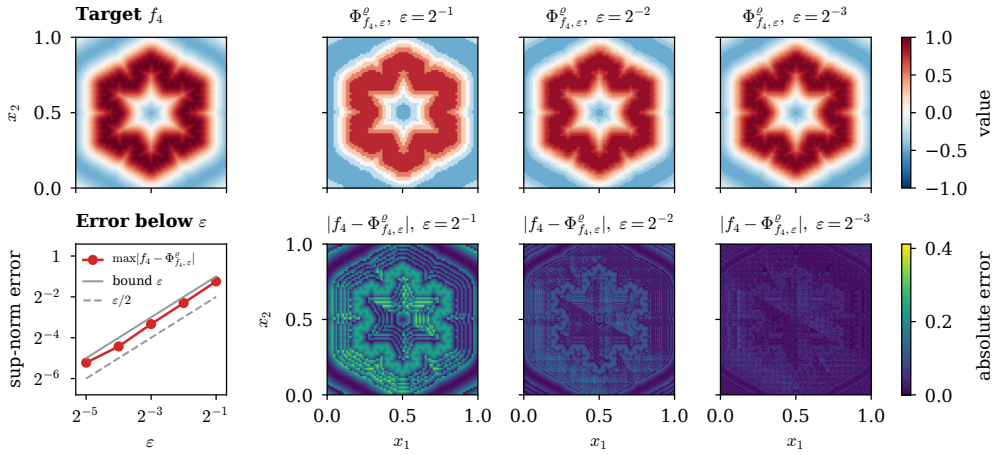


Figure 4: Numerical verification of Theorem 3.1 for the approximation of f_4 by $\Phi^{\varrho}_{f_4,\varepsilon}$. Top: The target function f_4 and the network outputs for $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}$ ($M = 74, 147, 293$). Bottom: The sup-norm error over the 3,004 test points versus ε, and the absolute errors $|f_4 - \Phi^{\varrho}_{f_4,\varepsilon}|$ on a 161^2 grid. The error plot additionally includes $\varepsilon = 2^{-4}, 2^{-5}$.

sequence of tested accuracies. The visualizations for f_1, f_2, f_3, f_5, and f_6 are provided in Appendix F (Figures 8–13). For the two-dimensional targets $f_1, \ldots, f_4$, the maximal-error plots are included in their respective visualization figures; for the three-dimensional targets f_5 and f_6, they are presented separately in Figure 13, together with those for $\Phi^{\sigma}_{f,\varepsilon}$ and $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$ in Theorem 4.1. All observed errors are below ε as guaranteed, and both $\mathrm{bit}(N_f)/(n\log_2(n+1))$ (for $\Phi^{\varrho}_{f,\varepsilon}$) and $\mathrm{bit}(A_f)/(K\log_2 B)$ (for $\Phi^{\sigma}_{f,\varepsilon}$ and $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$) approach 1 as ε decreases in every case, consistently with Theorem 5.1.

Implementation. The dyadic quantities and nonintegral rational parameters are stored exactly as `Fraction` objects; R, M, B, K, n and the integer parameters as Python integers; $(q_\ell)_{\ell=0}^{K-1}$ as a temporary `int64` array; and the large integers S_n, J_f, N_f as arbitrary-precision `mpz` objects. Arbitrary-precision arithmetic is necessary because $\mathrm{bit}(N_f)$ reaches the order of 10^7, far beyond the precision of double precision.

Now we discuss how to deal with the bulky lookup table definition of ϱ in the verification. As shown in (13), the first hidden layer uses only the negative branch, which is evaluated directly using the floor function. In the second hidden layer (14), since $t = N_f + \ell = S_n + nJ_f + \ell$, dividing $t - S_n$ by n yields quotient J_f and remainder ℓ. The activation ϱ then extracts the corresponding base-$(n+1)$ digit of J_f: $\varrho(t) = \lfloor \frac{J_f}{(n+1)^{n-1-\ell}} \rfloor - (n+1)\lfloor \frac{J_f}{(n+1)^{n-\ell}} \rfloor = q_\ell$. Although constructing N_f requires the values $(q_\ell)_{\ell=0}^{K-1}$, once N_f has been formed, pointwise network evaluation recovers each required q_ℓ directly from N_f, without repeatedly accessing a lookup table.

7 Discussion

Limitations. Our results concern representational capacity rather than practical training. The encoding integers N_f and A_f may require $\Theta(\varepsilon^{-d/\alpha}\log(1/\varepsilon))$ bits, far exceeding standard fixed-precision floating-point arithmetic in the gradient descent-based training in actual function fitting. Moreover, in the $(d, 1)$ construction, ϱ is a lookup function rather than a practical nonlinearity, and its evaluation can be computationally expensive.

Open problems. (i) Can the optimal count $d+1$ be attained with an elementary discontinuous activation, or can $d+2$ neurons suffice without a skip connection? (ii) What is the minimal neuron count for uniform approximation with a continuous activation since $d+1$ neurons do not suffice when $d = 2$? (iii) Can a fixed architecture exploit the bounded differences between neighboring quantized values to attain bit complexity $\Theta(\varepsilon^{-d/\alpha})$, removing the logarithmic factor in our constructions?

AI USE STATEMENT

The research problem, main network constructions, theorem statements, and overall proof framework presented in this work were developed by the authors. A generative AI tool suggested the initial form of $f_h(\boldsymbol{x})$ used in the proof of Proposition 3.3 and an initial form of $f_{\mathrm{c}}(\boldsymbol{x})$ used in the proof of Proposition E.1.

The initial version of the exact-arithmetic verification code released as supplementary material was written by the authors. Generative AI tools subsequently helped refine and reorganize the scripts, remove development comments, and draft the README.

Generative AI tools also assisted with the presentation of mathematical arguments, language editing, manuscript organization, and summaries of literature selected and provided by the authors. The authors independently verified all AI-assisted arguments, literature summaries, and code and take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

Complete proofs of all theorems and of Propositions 3.2 and 3.3, which Theorem 3.4 relies on, are given in Appendices A–E. The activations ϱ and σ and the networks of Theorems 3.1 and 4.1 are given in closed form by (10), (1), (5), (12), and (16), which allows them to be evaluated exactly in Python; Section 6 reports this verification for the six test functions $f_1, \ldots, f_6$, and the code (three Python files and a README) is provided as supplementary material.

REFERENCES

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Namig J Guliyev and Vugar E Ismailov. A single hidden layer feedforward network with only one neuron in the hidden layer can approximate any univariate function. *Neural computation*, 28(7): 1289–1304, 2016.

Namig J Guliyev and Vugar E Ismailov. Approximation capability of two hidden layer feedforward neural networks with fixed weights. *Neurocomputing*, 316:262–269, 2018.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Vugar E Ismailov. On the approximation by neural networks with bounded number of neurons in hidden layers. *Journal of Mathematical Analysis and Applications*, 417(2):963–969, 2014.

HV Koch. Sur une courbe continue sans tangente, obtenue par une construction géométrique élémentaire. *Arkiv for Matematik, Astronomi och Fysik*, 1:681–704, 1904.

Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Vladimir Ya Lin and Allan Pinkus. Fundamentality of ridge functions. *Journal of Approximation Theory*, 75(3):295–311, 1993.

Vitaly Maiorov and Allan Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1-3):81–91, 1999.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 141:160–173, 2021.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on Learning Theory*, pp. 639–649. PMLR, 2018.

Dmitry Yarotsky. Elementary superexpressive activations. In *International Conference on Machine Learning*, pp. 11932–11940. PMLR, 2021.

Shijun Zhang, Zuowei Shen, and Haizhao Yang. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *Journal of Machine Learning Research*, 23(276):1–60, 2022.

APPENDIX

A DETAILS OF SECTION 3.1

The activation ϱ is well defined. Fix $n \in \mathbb{N}_+$. The unit intervals $I_{n,j,r}$ of (9), indexed by integers j and r such that $0 \le j < (n+1)^n$ and $0 \le r < n$, form a partition of $[S_n, S_{n+1})$. Indeed, let $k = nj + r$ ranging over $\{0, \dots, n(n+1)^n - 1\}$,

$$\bigcup_{j=0}^{(n+1)^n-1} \bigcup_{r=0}^{n-1} I_{n,j,r} = \bigcup_{k=0}^{n(n+1)^n-1} [S_n + k, S_n + k + 1) = [S_n, S_n + n(n+1)^n) = [S_n, S_{n+1}).$$

Since $S_1 = 0$ and $S_n \to \infty$, it follows that $\bigcup_{n=1}^{\infty} \bigcup_{j=0}^{(n+1)^n-1} \bigcup_{r=0}^{n-1} I_{n,j,r} = [0, \infty)$. Therefore, ϱ in (10) is well defined on $\mathbb{R}$. It is constant on each unit interval, takes values in $\{0, \dots, n\}$ on $[S_n, S_{n+1})$, and every compact set meets finitely many blocks; hence ϱ is locally bounded and Borel measurable, so $\varrho \in L^1_{\mathrm{loc}}(\mathbb{R})$.

Digit extraction identity of the activation ϱ. Write $j \in \{0, \dots, (n+1)^n - 1\}$ with base-$(n+1)$ expansion $j = \sum_{s=0}^{n-1} b_s (n+1)^{n-1-s}$, $b_s \in \{0, \dots, n\}$. For $r \in \{0, \dots, n-1\}$,

$$\Big\lfloor \frac{j}{(n+1)^{n-1-r}} \Big\rfloor = \sum_{s=0}^{r} b_s (n+1)^{r-s}, \quad \Big\lfloor \frac{j}{(n+1)^{n-r}} \Big\rfloor = \sum_{s=0}^{r-1} b_s (n+1)^{r-1-s},$$

Subtracting $n+1$ times the second from the first leaves b_r, i.e. $\lfloor j/(n+1)^{n-1-r} \rfloor - (n+1)\lfloor j/(n+1)^{n-r} \rfloor = b_r$. Since $S_n + nj + r \in I_{n,j,r}$ as defined in (9), this proves (11).

Parameters of the network $\Phi^{\varrho}_{f,\varepsilon}$. For the record, the parameters of $\Phi^{\varrho}_{f,\varepsilon}$ are: $W_1 = M \,\mathrm{Id}_d$, $\boldsymbol{b}_1 = -(M+1)\mathbf{1}$; $W_2 = \big(1, M+1, \dots, (M+1)^{d-1}\big)$, $b_2 = N_f + (M+1)\sum_{j=1}^{d}(M+1)^{j-1} = N_f + (M+1)\frac{(M+1)^d - 1}{M}$; $a = \delta$, $c = -R$. All are integers except δ (dyadic) and $-R$ (an integer once $R \in \mathbb{N}_+$).

B PROOFS OF SECTION 3.2

B.1 PROOF OF PROPOSITION 3.2

Define $f_0 : [0,1]^d \to \mathbb{R}$ by $f_0(\boldsymbol{x}) := \zeta \|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty^\alpha$, $\boldsymbol{x}_0 := \big(\frac{1}{2}, \dots, \frac{1}{2}\big)$, $\zeta := \min\{\lambda, 2^\alpha R\}$.

Step 1: Verification that $f_0 \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$. Since $\max_{\boldsymbol{x}\in[0,1]^d} \|\boldsymbol{x}-\boldsymbol{x}_0\|_\infty = \frac{1}{2}$, we have $\|f_0\|_{L^\infty([0,1]^d)} = \zeta 2^{-\alpha} \le R$. For $a, b \ge 0$ and $\alpha \in (0,1]$, one has $|a^\alpha - b^\alpha| \le |a-b|^\alpha$ by the subadditivity of $t \mapsto t^\alpha$. Hence, by the reverse triangle inequality,

$$\begin{aligned}|f_0(\boldsymbol{x}) - f_0(\boldsymbol{y})| &= \zeta\big|\|\boldsymbol{x}-\boldsymbol{x}_0\|_\infty^\alpha - \|\boldsymbol{y}-\boldsymbol{x}_0\|_\infty^\alpha\big| \le \zeta\big|\|\boldsymbol{x}-\boldsymbol{x}_0\|_\infty - \|\boldsymbol{y}-\boldsymbol{x}_0\|_\infty\big|^\alpha \\ &\le \zeta\|\boldsymbol{x}-\boldsymbol{y}\|_\infty^\alpha \le \lambda\|\boldsymbol{x}-\boldsymbol{y}\|_\infty^\alpha.\end{aligned}$$

Step 2: Error lower bound from the kernel of the first layer. The input enters Φ only through $W_1\boldsymbol{x}+\boldsymbol{b}_1$ with $W_1 \in \mathbb{R}^{n_1\times d}$. Since $n_1 < d$, there exists $\boldsymbol{z} \in \ker W_1$ with $\|\boldsymbol{z}\|_\infty = 1$; put $\boldsymbol{x}_1 := \boldsymbol{x}_0 + \frac{1}{2}\boldsymbol{z} \in [0,1]^d$. Then $W_1\boldsymbol{x}_1 + \boldsymbol{b}_1 = W_1\boldsymbol{x}_0 + \boldsymbol{b}_1$, hence $\Phi(\boldsymbol{x}_1) = \Phi(\boldsymbol{x}_0)$, whereas $f_0(\boldsymbol{x}_0) = 0$ and $f_0(\boldsymbol{x}_1) = \zeta 2^{-\alpha}$. Therefore

$$\zeta 2^{-\alpha} = |f_0(\boldsymbol{x}_1) - f_0(\boldsymbol{x}_0)| \le |f_0(\boldsymbol{x}_1) - \Phi(\boldsymbol{x}_1)| + |\Phi(\boldsymbol{x}_0) - f_0(\boldsymbol{x}_0)| \le 2\|f_0 - \Phi\|_{L^\infty([0,1]^d)}.$$

Therefore, by definition of ζ, $\|f_0-\Phi\|_{L^\infty([0,1]^d)} \ge \frac{1}{2}\zeta 2^{-\alpha} = \frac{1}{2}\min\{\lambda 2^{-\alpha}, R\} =: \varepsilon_0 > 0$. □

B.2 Proof of Proposition 3.3

Let $\Omega := (0,1)^d$ and define $h : \mathbb{R}^d \to \mathbb{R}$, $h(\boldsymbol{x}) := \exp(\|\boldsymbol{x}\|_2^2)$. We write $(C_c^\infty(\Omega))'$ for the continuous dual of $C_c^\infty(\Omega)$.

Step 1: a suitable directional derivative of Φ vanishes. Let $\Phi(\boldsymbol{x}) = c + \sum_{i=1}^{n_1} a_i\phi(\boldsymbol{w}_i\cdot\boldsymbol{x}+b_i) \in \mathcal{N}((n_1);\phi)$. Since $\phi \in L^1_{\mathrm{loc}}(\mathbb{R})$, each function $a_i\phi(\boldsymbol{w}_i\cdot\boldsymbol{x}+b_i)$ is locally integrable on $\mathbb{R}^d$ and can therefore be viewed as a regular distribution on Ω.

Denote $\mathbb{S}^{d-1} := \{\boldsymbol{x}\in\mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$. For each $i = 1,\dots,n_1$, choose a unit vector $\boldsymbol{v}_i \in \mathbb{S}^{d-1}$ satisfying $\boldsymbol{v}_i\cdot\boldsymbol{w}_i = 0$. Such a vector exists because $d \ge 2$; if $\boldsymbol{w}_i = 0$, we may choose any unit vector. Write $D_{\boldsymbol{v}} := \boldsymbol{v}\cdot\nabla$ for the distributional directional derivative.

For every test function $\psi \in C_c^\infty(\Omega)$, extended by zero to $\mathbb{R}^d$, we have $\big\langle D_{\boldsymbol{v}_i}\big[a_i\phi(\boldsymbol{w}_i\cdot\boldsymbol{x}+b_i)\big], \psi\big\rangle = -\int_{\mathbb{R}^d} a_i\phi(\boldsymbol{w}_i\cdot\boldsymbol{x}+b_i) D_{\boldsymbol{v}_i}\psi(\boldsymbol{x})\,d\boldsymbol{x}$. Using the orthogonal decomposition $\boldsymbol{x} = \boldsymbol{y} + s\boldsymbol{v}_i$, $\boldsymbol{y} \in \boldsymbol{v}_i^\perp$, whose Jacobian is one, and the identity $\boldsymbol{w}_i\cdot(\boldsymbol{y}+s\boldsymbol{v}_i) = \boldsymbol{w}_i\cdot\boldsymbol{y}$ gives

$$\big\langle D_{\boldsymbol{v}_i}\big[a_i\phi(\boldsymbol{w}_i\cdot\boldsymbol{x}+b_i)\big], \psi\big\rangle = -\int_{\boldsymbol{v}_i^\perp} a_i\phi(\boldsymbol{w}_i\cdot\boldsymbol{y}+b_i)\Big[\int_{\mathbb{R}} \frac{\partial}{\partial s}\psi(\boldsymbol{y}+s\boldsymbol{v}_i)\,ds\Big]d\boldsymbol{y} = 0.$$

The inner integral vanishes because ψ has compact support. Therefore,

$$D_{\boldsymbol{v}_i}\big[a_i\phi(\boldsymbol{w}_i\cdot\boldsymbol{x}+b_i)\big] = 0 \quad \text{in } (C_c^\infty(\Omega))'. \tag{22}$$

Now define $\mathcal{L}_{\boldsymbol{v}} := D_{\boldsymbol{v}_1}D_{\boldsymbol{v}_2}\cdots D_{\boldsymbol{v}_{n_1}}$, $\boldsymbol{v} := (\boldsymbol{v}_1,\dots,\boldsymbol{v}_{n_1}) \in (\mathbb{S}^{d-1})^{n_1}$. Constant-coefficient directional derivatives commute. For each i, the operator $\mathcal{L}_{\boldsymbol{v}}$ contains the factor $D_{\boldsymbol{v}_i}$, which makes the ith ridge term vanish by (22). It also makes the constant term vanish. Hence $\mathcal{L}_{\boldsymbol{v}}\Phi = 0$ in $(C_c^\infty(\Omega))'$.

Step 2: the same directional derivative of h does not vanish. For every $\boldsymbol{u} \in \mathbb{R}^d$, $D_{\boldsymbol{u}}h(\boldsymbol{x}) = 2(\boldsymbol{u}\cdot\boldsymbol{x})h(\boldsymbol{x})$. Repeated application of the product rule gives $\mathcal{L}_{\boldsymbol{v}}h = P_{\boldsymbol{v}}h$, where $P_{\boldsymbol{v}}$ is a polynomial of degree n_1 whose homogeneous part of degree n_1 is $2^{n_1}\prod_{i=1}^{n_1}(\boldsymbol{v}_i\cdot\boldsymbol{x})$. Since each $\boldsymbol{v}_i$ is a unit vector, $\boldsymbol{v}_i\cdot\boldsymbol{x}$ is a nonzero linear polynomial. Therefore, $\prod_{i=1}^{n_1}(\boldsymbol{v}_i\cdot\boldsymbol{x})$ is a nonzero polynomial. It follows that $P_{\boldsymbol{v}} \not\equiv 0$, and hence

$$\mathcal{L}_{\boldsymbol{v}}h = P_{\boldsymbol{v}}h \not\equiv 0 \tag{23}$$

for every $\boldsymbol{v} \in (\mathbb{S}^{d-1})^{n_1}$.

Step 3: derive a uniform lower bound for the approximation error. Choose $\chi \in C_c^\infty(\Omega)$ such that $\chi \ge 0$ and $\chi > 0$ on a nonempty open subset of Ω, and define $A_{n_1}(\boldsymbol{v}) := \int_\Omega \chi(\boldsymbol{x})\big|\mathcal{L}_{\boldsymbol{v}}h(\boldsymbol{x})\big|^2\,d\boldsymbol{x}$. The function $\mathcal{L}_{\boldsymbol{v}}h$ is real analytic and, by (23), is not identically zero. It therefore cannot vanish on the open set where $\chi > 0$, and hence $A_{n_1}(\boldsymbol{v}) > 0$. Moreover, $\boldsymbol{v} \mapsto A_{n_1}(\boldsymbol{v})$ is continuous. Since $(\mathbb{S}^{d-1})^{n_1}$ is compact,

$$a_{n_1} := \min_{\boldsymbol{v}\in(\mathbb{S}^{d-1})^{n_1}} A_{n_1}(\boldsymbol{v}) > 0. \tag{24}$$

For the directions $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n_1}$ selected in Step 1, define $\psi_{\boldsymbol{v}} := \chi\, \mathcal{L}_{\boldsymbol{v}} h \in C_c^\infty(\Omega)$. Since $\mathcal{L}_{\boldsymbol{v}}\Phi = 0$ in $(C_c^\infty(\Omega))'$, we obtain $A_{n_1}(\boldsymbol{v}) = \langle \mathcal{L}_{\boldsymbol{v}} h, \psi_{\boldsymbol{v}} \rangle = \langle \mathcal{L}_{\boldsymbol{v}}(h-\Phi), \psi_{\boldsymbol{v}} \rangle = (-1)^{n_1} \int_\Omega (h-\Phi) \mathcal{L}_{\boldsymbol{v}} \psi_{\boldsymbol{v}}\, d\boldsymbol{x}$. Therefore,

$$A_{n_1}(\boldsymbol{v}) \le \|h - \Phi\|_{L^\infty(\Omega)} \|\mathcal{L}_{\boldsymbol{v}} \psi_{\boldsymbol{v}}\|_{L^1(\Omega)}. \tag{25}$$

By the Leibniz rule, $\mathcal{L}_{\boldsymbol{v}} \psi_{\boldsymbol{v}} = \sum_{S \subseteq \{1,\dots,n_1\}} \big(\prod_{i \in S} D_{\boldsymbol{v}_i}\big)\chi \cdot \big(\prod_{i \notin S} D_{\boldsymbol{v}_i}\big) \mathcal{L}_{\boldsymbol{v}} h$ involves derivatives of the fixed smooth functions χ and h of order at most $2n_1$ with coefficients depending polynomially in $\boldsymbol{v}$. Hence $(\boldsymbol{x}, \boldsymbol{v}) \mapsto \mathcal{L}_{\boldsymbol{v}} \psi_{\boldsymbol{v}}(\boldsymbol{x})$ is continuous on the compact set $\operatorname{supp} \chi \times (\mathbb{S}^{d-1})^{n_1}$. It follows that $0 < b_{n_1} := \sup_{\boldsymbol{v} \in (\mathbb{S}^{d-1})^{n_1}} \|\mathcal{L}_{\boldsymbol{v}} \psi_{\boldsymbol{v}}\|_{L^1(\Omega)} < \infty$. Combining (25) with (24) gives

$$\|h - \Phi\|_{L^\infty(\Omega)} \ge \frac{A_{n_1}(\boldsymbol{v})}{b_{n_1}} \ge \frac{a_{n_1}}{b_{n_1}} > 0, \quad \text{for every } \Phi \in \mathcal{N}((n_1); \phi). \tag{26}$$

Step 4: rescale h into the Hölder class. On $[0,1]^d$, $\|h\|_\infty = e^d$ and $\|\nabla h(\boldsymbol{x})\|_1 = 2e^{\|\boldsymbol{x}\|_2^2} \sum_{j=1}^d x_j \le 2de^d$. Therefore, by the mean value theorem, $|h(\boldsymbol{x}) - h(\boldsymbol{y})| \le 2de^d \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \le 2de^d \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^\alpha$, where the last inequality follows from $\|\boldsymbol{x} - \boldsymbol{y}\|_\infty \le 1$ and $\alpha \le 1$.

Set $\gamma := \min\big\{Re^{-d}, \frac{\lambda}{2de^d}\big\}$, $f_h := \gamma h$. Then $f_h \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$. Since $\mathcal{N}((n_1); \phi)$ is invariant under multiplication by the positive constant γ, it follows from (26) that

$$\inf_{\Phi \in \mathcal{N}((n_1);\phi)} \|f_h - \Phi\|_{L^\infty([0,1]^d)} = \gamma \inf_{\Phi \in \mathcal{N}((n_1);\phi)} \|h - \Phi\|_{L^\infty([0,1]^d)} \ge \gamma \frac{a_{n_1}}{b_{n_1}} =: \varepsilon_1 > 0.$$

□

B.3 Proof of Theorem 3.4

Suppose, to the contrary, that $|\boldsymbol{n}| \le d$. If $L \ge 2$, then $n_1 \le |\boldsymbol{n}| - n_2 \le d - 1 < d$, so Proposition 3.2 exhibits an $f_0 \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$ that cannot be approximated by any $\Phi \in \mathcal{N}^{\mathrm{skip}}(\boldsymbol{n}; \phi)$ with error smaller than ε_0. If $L = 1$, then $\mathcal{N}^{\mathrm{skip}}(\boldsymbol{n}; \phi) = \mathcal{N}((n_1); \phi)$, and Proposition 3.3 exhibits an $f_h \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$ that cannot be approximated by any such network with error smaller than ε_1. Both cases contradict arbitrary-accuracy approximation. Therefore, $|\boldsymbol{n}| \ge d + 1$.

Together with the upper bound provided by Theorem 3.1, this yields $\nu(\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)) = d + 1$. □

C Details and Proofs of Section 4

Parameters of the networks $\Phi^\sigma_{f,\varepsilon}$ and $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$. $\Phi^\sigma_{f,\varepsilon}$: $W_1 = M \operatorname{Id}_d$, $\boldsymbol{b}_1 = 0$; $W_2 = -\log_2(B)\, \boldsymbol{c}^\top$, $b_2 = -1$; $W_3 = (2A_f,\, 2A_f/B)^\top$, $\boldsymbol{b}_3 = (2A_f,\, 2A_f/B)^\top$; $\boldsymbol{a} = (\delta, -\delta B)$, $c = -R$.

$\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$: the same first two layers; $W_3 = b_3 = 2A_f/B$; output weights $2\delta A_f$ on v and $-\delta B$ on w^{skip}, bias $-R + 2\delta A_f$. All parameters are integers or dyadic rationals.

C.1 Proof of Theorem 4.1

Proof. For $\Phi^\sigma_{f,\varepsilon}$, (20) and (7) give $|f(\boldsymbol{x}) - \Phi^\sigma_{f,\varepsilon}(\boldsymbol{x})| < \varepsilon$ on every cell.

For $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$, on Q_ℓ we have $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon} = \widehat{f}_\ell + \delta\eta_\ell$ with $0 \le \delta\eta_\ell < \delta$ and $0 \le f(\boldsymbol{x}^{(\ell)}) - \widehat{f}_\ell < \delta$ by (6), so $|f(\boldsymbol{x}^{(\ell)}) - \Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}(\boldsymbol{x})| = |(f(\boldsymbol{x}^{(\ell)}) - \widehat{f}_\ell) - \delta\eta_\ell| < \delta$ and the triangle inequality with $|f(\boldsymbol{x}) - f(\boldsymbol{x}^{(\ell)})| \le \lambda M^{-\alpha} \le \varepsilon/2$ gives the claim. □

D Proof of Theorem 5.1

Throughout the proof, we consider $d \ge 2$, $\alpha \in (0, 1]$, $\lambda > 0$, $R \in \mathbb{N}_+$ are fixed and $\varepsilon \to 0$; by (8), $M = \Theta(\varepsilon^{-1/\alpha})$, $K = \Theta(\varepsilon^{-d/\alpha})$, $B = \Theta(\varepsilon^{-1})$, and $\log_2 B = \Theta(\log(1/\varepsilon))$.

D.1 Bit complexity of A_f

Upper bound. Since $0 \le q_\ell \le B-1$, $0 \le A_f = \sum_{\ell=0}^{K-1} q_\ell B^\ell \le (B-1)\sum_{\ell=0}^{K-1} B^\ell = B^K - 1$, hence $\mathrm{bit}(A_f) \le \lceil \log_2 B^K \rceil = K \log_2 B$ (an integer, as B is a power of two). Thus $\sup_{f \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}(A_f) = O(\varepsilon^{-d/\alpha}\log(1/\varepsilon))$.

Lower bound. Let $f_R \equiv R$. By (5), $B/2 < \lfloor 2R/\delta \rfloor + 1 \le B$ as soon as $\lfloor 2R/\delta \rfloor + 1 \ge 2$, i.e. for small ε; since $B/2$ and $\lfloor 2R/\delta \rfloor$ are integers, $B/2 \le \lfloor 2R/\delta \rfloor \le B-1$. For f_R every digit equals $q_\ell = \lfloor 2R/\delta \rfloor \ge B/2$, so

$$A_{f_R} \ge \frac{B}{2}\sum_{\ell<K} B^\ell \ge \frac{1}{2}B^K$$

and $\mathrm{bit}(A_{f_R}) \ge K\log_2 B - 1$. Hence

$$\sup_{f \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}(A_f) = \Theta\big(K\log_2 B\big) = \Theta\big(\varepsilon^{-d/\alpha}\log\frac{1}{\varepsilon}\big).$$

D.2 Bit complexity of $\Phi^\sigma_{f,\varepsilon}$ and $\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}$

Using the parameter lists in Appendix C: the first hidden layer has d nonzero weights equal to M, contributing $d\ \mathrm{bit}(M) = O(d\log(1/\varepsilon))$. The second hidden layer has weights $\log_2(B)(M+1)^{j-1}$, $j = 1,\dots,d$, and bias -1, with $\mathrm{bit}(\log_2(B)(M+1)^{j-1}) = O(j\log(1/\varepsilon))$, contributing $O(d^2\log(1/\varepsilon))$ in total. The third hidden layer contains $2A_f$ and $2A_f/B$ (each at most twice); since $2A_f/B$ in lowest terms has numerator at most $2A_f$ and denominator at most B, $\mathrm{bit}(2A_f/B) \le \mathrm{bit}(A_f) + \mathrm{bit}(B) + O(1)$, so this layer contributes $O(\mathrm{bit}(A_f) + \log(1/\varepsilon))$. The output layer has parameters δ, δB, R (and $2\delta A_f$, $-R+2\delta A_f$ in the skip variant), contributing $O(\mathrm{bit}(A_f)+\log(1/\varepsilon))$. Altogether, since $d^2\log(1/\varepsilon) = o\big(\varepsilon^{-d/\alpha}\log(1/\varepsilon)\big)$ as $\varepsilon \to 0$, $\sup_{f\in\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(\Phi^\sigma_{f,\varepsilon}\big) = \Theta(\sup_{f\in\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(A_f\big))$ and $\sup_{f\in\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}\big) = \Theta(\sup_{f\in\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(A_f\big))$. Thus,

$$\sup_{f\in\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(\Phi^\sigma_{f,\varepsilon}\big) = \Theta\big(\varepsilon^{-d/\alpha}\log\frac{1}{\varepsilon}\big),\quad \sup_{f\in\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)} \mathrm{bit}\big(\Phi^{\sigma,\mathrm{skip}}_{f,\varepsilon}\big) = \Theta\big(\varepsilon^{-d/\alpha}\log\frac{1}{\varepsilon}\big).$$

D.3 Bit complexity of $\Phi^\varrho_{f,\varepsilon}$

By the parameter list at Appendix A, all parameters other than the bias $b_2 = N_f + (M+1)((M+1)^d - 1)/M$ contribute $O(d^2\log(1/\varepsilon))$ bits, and $\mathrm{bit}(b_2) = \mathrm{bit}(N_f) + O(d\log(1/\varepsilon))$. Since $n = \max\{K, B-1\}$ and $K = \Theta(\varepsilon^{-d/\alpha})$ dominates $B = \Theta(\varepsilon^{-1})$ (as $d/\alpha \ge 2$), we have $n = K = \Theta(\varepsilon^{-d/\alpha})$ for small ε. From $S_n \le N_f < S_{n+1}$ and

$$(n-1)n^{\,n-1} \le S_n = \sum_{k=1}^{n-1} k(k+1)^k \le (n-1)\cdot(n-1)n^{\,n-1} \le n^{\,n+1},\quad S_{n+1} \le (n+1)^{\,n+2},$$

we get $(n-1)\log_2 n \le \log_2 N_f \le (n+2)\log_2(n+1)$, i.e. $\mathrm{bit}(N_f) = \Theta(n\log n) = \Theta(\varepsilon^{-d/\alpha}\log(1/\varepsilon))$ for every $f \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$. Hence $\sup_f \mathrm{bit}(\Phi^\varrho_{f,\varepsilon}) = \Theta(\varepsilon^{-d/\alpha}\log(1/\varepsilon))$. □

E Continuous activations: uniform obstruction and L^p construction

The activation ϱ of Theorem 3.1 is discontinuous, while Theorem 3.4 allows arbitrary locally integrable activations. Appendix E.1 shows that the discontinuity is not incidental: with a continuous activation, $d+1$ hidden neurons do not suffice for uniform approximation, already for $d = 2$. Appendix E.2 shows that the obstruction is specific to the uniform norm: a single continuous activation ϱ_c, independent of f, ε, p and d, gives the $(d,1)$ architecture an L^p guarantee for every $p \in [1,\infty)$.

E.1 A UNIFORM-APPROXIMATION OBSTRUCTION FOR $(2,1)$ NETWORKS

Proposition E.1 (Three-neuron obstruction for continuous activations)**.** *Let $d = 2$, $\alpha \in (0,1]$, $\lambda, R > 0$, and let $\phi : \mathbb{R} \to \mathbb{R}$ be continuous. For every $L \in \mathbb{N}_+$ and every architecture $\boldsymbol{n} = (n_1, \dots, n_L) \in \mathbb{N}_+^L$ with $|\boldsymbol{n}| = 3$, there exist a function $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ and a constant $\varepsilon_2 > 0$, depending only on α, λ, R, such that*

$$\inf_{\Phi \in \mathcal{N}(\boldsymbol{n};\phi)} \|f - \Phi\|_{L^\infty([0,1]^d)} \geq \varepsilon_2.$$

Consequently, no fixed architecture with three hidden neurons and a continuous activation can approximate $\mathcal{H}^{\alpha}_{\lambda,R}([0,1]^d)$ to arbitrary accuracy.

Proof. For $|\boldsymbol{n}| = 3$, the only possible architectures are (3), $(1,2)$, $(2,1)$, $(1,1,1)$. The architecture (3) is excluded by Proposition 3.3, while $(1,2)$ and $(1,1,1)$ are excluded by Proposition 3.2. It therefore remains to only consider $\boldsymbol{n} = (2,1)$.

Step 0: growth of the function and the collision condition. Let $\boldsymbol{x}_0 := (\frac{1}{2}, \frac{1}{2})$. For $\boldsymbol{x} \in [0,1]^2$, write $\boldsymbol{y} := \boldsymbol{x} - \boldsymbol{x}_0 = (y_1, y_2)$, and define

$$f_{\mathrm{c}}(\boldsymbol{x}) := \|\boldsymbol{y}\|_2 + \frac{1}{10}\kappa(\boldsymbol{y}), \quad \kappa(\boldsymbol{y}) := y_1^3 + 2y_2^3. \tag{27}$$

Since $\boldsymbol{y} \in [-\frac{1}{2}, \frac{1}{2}]^2$, we have $|\kappa(\boldsymbol{y})| \leq |y_1|^3 + 2|y_2|^3 \leq \frac{3}{4}\|\boldsymbol{y}\|_2$. Thus, with $\theta := 3/40$,

$$(1-\theta)\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq f_{\mathrm{c}}(\boldsymbol{x}) \leq (1+\theta)\|\boldsymbol{x} - \boldsymbol{x}_0\|_2. \tag{28}$$

In particular, $f_{\mathrm{c}}(\boldsymbol{x}_0) = 0$.

Let $\Phi \in \mathcal{N}((2,1);\phi)$. In the notation of (3), write $\boldsymbol{\eta}^\top, \boldsymbol{\xi}^\top \in \mathbb{R}^2$ for the rows of W_1, $\boldsymbol{b}_1 = (b_{1,1}, b_{1,2})$, $W_2 = (w_1, w_2)$, $\boldsymbol{b}_2 = b_2$, and $\boldsymbol{a} = a$. Then $\Phi = g \circ h$ with $h(\boldsymbol{x}) = \psi_1(\boldsymbol{\eta}^\top \boldsymbol{x}) + \psi_2(\boldsymbol{\xi}^\top \boldsymbol{x})$, where $\psi_i(t) := w_i\,\phi(t + b_{1,i})$ and $g(t) := a\,\phi(t + b_2) + c$; thus $g, \psi_1, \psi_2 : \mathbb{R} \to \mathbb{R}$ are continuous. Suppose that $\|f_{\mathrm{c}} - \Phi\|_{L^\infty([0,1]^2)} \leq \varepsilon$. Whenever $h(\boldsymbol{x}) = h(\boldsymbol{x}')$, we have $\Phi(\boldsymbol{x}) = g(h(\boldsymbol{x})) = g(h(\boldsymbol{x}')) = \Phi(\boldsymbol{x}')$, and hence, for $\boldsymbol{x}, \boldsymbol{x}' \in [0,1]^2$, the *collision condition*,

$$h(\boldsymbol{x}) = h(\boldsymbol{x}') \implies |f_{\mathrm{c}}(\boldsymbol{x}) - f_{\mathrm{c}}(\boldsymbol{x}')| \leq |f_{\mathrm{c}}(\boldsymbol{x}) - \Phi(\boldsymbol{x})| + |\Phi(\boldsymbol{x}') - f_{\mathrm{c}}(\boldsymbol{x}')| \leq 2\varepsilon. \tag{29}$$

Steps 1–4 then show that (29) forces $\varepsilon \geq \varepsilon_*$ for an absolute constant $\varepsilon_* > 0$ defined in Step 4.

Step 1: a direct error bound for the degenerate case $\boldsymbol{\eta} \parallel \boldsymbol{\xi}$. Suppose that $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are linearly dependent. There is then a unit vector $\boldsymbol{v}$ orthogonal to both of them, so that $h(\boldsymbol{x}_0 + t\boldsymbol{v}) = h(\boldsymbol{x}_0)$ whenever $\boldsymbol{x}_0 + t\boldsymbol{v} \in [0,1]^2$. This line segment reaches the boundary of the square and therefore contains a point $\boldsymbol{p} \in \partial[0,1]^2$ satisfying $\|\boldsymbol{p} - \boldsymbol{x}_0\|_2 \geq \frac{1}{2}$. By (29) and (28), $2\varepsilon \geq |f_{\mathrm{c}}(\boldsymbol{p}) - f_{\mathrm{c}}(\boldsymbol{x}_0)| = f_{\mathrm{c}}(\boldsymbol{p}) \geq (1-\theta)\|\boldsymbol{p} - \boldsymbol{x}_0\|_2 \geq \frac{1-\theta}{2}$. Thus $\varepsilon \geq \frac{1-\theta}{4}$.

We may henceforth assume that $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are linearly independent.

Step 2: coordinates adapted to the two directions. Absorbing $\|\boldsymbol{\eta}\|_2$ and $\|\boldsymbol{\xi}\|_2$ into ψ_1 and ψ_2, we may assume $\|\boldsymbol{\eta}\|_2 = \|\boldsymbol{\xi}\|_2 = 1$. Put $\Delta := |\det(\boldsymbol{\eta}, \boldsymbol{\xi})| \in (0,1]$ and let $\boldsymbol{e}$ be the unit vector orthogonal to $\boldsymbol{\xi}$ with $\boldsymbol{\eta}^\top \boldsymbol{e} = \Delta$, and $\boldsymbol{e}'$ the unit vector orthogonal to $\boldsymbol{\eta}$ with $\boldsymbol{\xi}^\top \boldsymbol{e}' = \Delta$. For $\boldsymbol{x} \in \mathbb{R}^2$ define $u(\boldsymbol{x}) := \frac{\boldsymbol{\eta}^\top(\boldsymbol{x} - \boldsymbol{x}_0)}{\Delta}$, $v(\boldsymbol{x}) := \frac{\boldsymbol{\xi}^\top(\boldsymbol{x} - \boldsymbol{x}_0)}{\Delta}$. Then $\boldsymbol{x} = \boldsymbol{x}_0 + u(\boldsymbol{x})\boldsymbol{e} + v(\boldsymbol{x})\boldsymbol{e}'$: applying $\boldsymbol{\eta}^\top$ and $\boldsymbol{\xi}^\top$ to both sides gives the same values, and $\boldsymbol{\eta}, \boldsymbol{\xi}$ span $\mathbb{R}^2$. In these coordinates $h(\boldsymbol{x}) = \psi_1(\boldsymbol{\eta}^\top \boldsymbol{x}_0 + \Delta u(\boldsymbol{x})) + \psi_2(\boldsymbol{\xi}^\top \boldsymbol{x}_0 + \Delta v(\boldsymbol{x}))$; absorbing the affine changes of variables into ψ_1, ψ_2 we may write

$$h(\boldsymbol{x}) = \psi_1\big(u(\boldsymbol{x})\big) + \psi_2\big(v(\boldsymbol{x})\big). \tag{30}$$

Step 3: constructing pairs with the same h-value. Let $L_0 := \{\boldsymbol{x} \in [0,1]^2 : h(\boldsymbol{x}) = h(\boldsymbol{x}_0)\}$. If L_0 contains a point $\boldsymbol{p}$ with $\|\boldsymbol{p} - \boldsymbol{x}_0\|_2 \geq \frac{1}{4}$, then (29) and (28) give $2\varepsilon \geq f_{\mathrm{c}}(\boldsymbol{p}) \geq \frac{1-\theta}{4}$, i.e. $\varepsilon \geq \frac{1-\theta}{8}$.

It remains to consider the case $L_0 \subset B(\boldsymbol{x}_0, \frac{1}{4})$, where $B(\boldsymbol{x}_0, \rho) := \{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 < \rho\}$ denotes the open Euclidean ball. Then the continuous function $h(\boldsymbol{x}) - h(\boldsymbol{x}_0)$ does not vanish on

$[0,1]^2 \setminus B(\boldsymbol{x}_0, \frac{1}{4})$, which is connected, and hence has constant sign there. Replacing (ψ_1, ψ_2, g) by $(-\psi_1, -\psi_2, g(-\cdot))$ if necessary, which leaves Φ unchanged, we may assume

$$h(\boldsymbol{x}) > h(\boldsymbol{x}_0) \quad \text{for all } \boldsymbol{x} \in [0,1]^2 \setminus B\big(\boldsymbol{x}_0, \tfrac{1}{4}\big). \tag{31}$$

The points $\boldsymbol{x}_0 \pm \frac{1}{4}\boldsymbol{e}$ have adapted coordinates $(\pm\frac{1}{4}, 0)$ and lie on $\partial B(\boldsymbol{x}_0, \frac{1}{4})$, hence in $[0,1]^2 \setminus B(\boldsymbol{x}_0, \frac{1}{4})$. By (31) and (30), $\psi_1(\pm\frac{1}{4}) + \psi_2(0) > \psi_1(0) + \psi_2(0)$, i.e. $\psi_1\big(-\frac{1}{4}\big) > \psi_1(0)$, $\psi_1\big(\frac{1}{4}\big) > \psi_1(0)$. Replacing $(\boldsymbol{\eta}, \boldsymbol{e}, \psi_1)$ by $(-\boldsymbol{\eta}, -\boldsymbol{e}, \psi_1(-\cdot))$ if necessary, which leaves h, Δ and (30) unchanged, we may assume $\psi_1(-\frac{1}{4}) \le \psi_1(\frac{1}{4})$. Since $\psi_1(0) < \psi_1(-\frac{1}{4}) \le \psi_1(\frac{1}{4})$, the intermediate value theorem on $[0, \frac{1}{4}]$ yields $r \in (0, \frac{1}{4}]$ with

$$\psi_1(r) = \psi_1\big(-\tfrac{1}{4}\big). \tag{32}$$

For $|t| \le \frac{1}{4}$ define $\boldsymbol{x}_1(t) := \boldsymbol{x}_0 - \frac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'$ and $\boldsymbol{x}_2(t) := \boldsymbol{x}_0 + r\boldsymbol{e} + t\boldsymbol{e}'$, with adapted coordinates $(-\frac{1}{4}, t)$ and (r, t). Both lie in the closed ball of radius $\frac{1}{4} + |t| \le \frac{1}{2}$ about $\boldsymbol{x}_0$, hence in $[0,1]^2$, and by (30) and (32), $h(\boldsymbol{x}_1(t)) = \psi_1(-\frac{1}{4}) + \psi_2(t) = \psi_1(r) + \psi_2(t) = h(\boldsymbol{x}_2(t))$. Consequently (29) gives

$$\big|f_{\mathrm{c}}(\boldsymbol{x}_1(t)) - f_{\mathrm{c}}(\boldsymbol{x}_2(t))\big| \le 2\varepsilon, \quad |t| \le \tfrac{1}{4}.$$

At $t = 0$, (28) gives $f_{\mathrm{c}}(\boldsymbol{x}_1(0)) \ge \frac{1-\theta}{4}$ and $f_{\mathrm{c}}(\boldsymbol{x}_2(0)) \le (1+\theta)r$, so $\frac{1-\theta}{4} - 2\varepsilon \le (1+\theta)r$. With $\theta = \frac{3}{40}$ this yields $r \ge \frac{1}{8}$ whenever $\varepsilon \le \frac{1}{50}$; in that case $r \in [\frac{1}{8}, \frac{1}{4}]$.

Step 4: a uniform gap between their f_{c}-values. Let $\mathbb{S}^1 := \{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_2 = 1\}$ and, on the compact set $\mathcal{K} := \mathbb{S}^1 \times \mathbb{S}^1 \times [\frac{1}{8}, \frac{1}{4}]$, define

$$\mathcal{M}(\boldsymbol{e}, \boldsymbol{e}', r) := \max_{|t| \le 1/4} \big|f_{\mathrm{c}}\big(\boldsymbol{x}_0 - \tfrac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'\big) - f_{\mathrm{c}}\big(\boldsymbol{x}_0 + r\boldsymbol{e} + t\boldsymbol{e}'\big)\big|.$$

$\mathcal{M}$ is continuous on $\mathcal{K}$. We claim that

$$\mathcal{M}(\boldsymbol{e}, \boldsymbol{e}', r) > 0 \quad \text{for every } (\boldsymbol{e}, \boldsymbol{e}', r) \in \mathcal{K}. \tag{33}$$

Note here we include parallel pairs of unit vectors in $\mathcal{K}$ (not only those produced by Step 2) so that the parameter set is compact and yields a uniform lower bound.

Case $\boldsymbol{e}' = \pm\boldsymbol{e}$. Choose $t = \pm\frac{1}{4}$ accordingly, so that $\boldsymbol{x}_0 - \frac{1}{4}\boldsymbol{e} + t\boldsymbol{e}' = \boldsymbol{x}_0$, while $\boldsymbol{x}_0 + r\boldsymbol{e} + t\boldsymbol{e}' = \boldsymbol{x}_0 + (r + \frac{1}{4})\boldsymbol{e}$. The first value of f_{c} is 0, the second is at least $(1-\theta)(r+\frac{1}{4}) > 0$ by (28).

Case $\boldsymbol{e}' \nparallel \boldsymbol{e}$. Put $\mu := \boldsymbol{e}^{\top}\boldsymbol{e}' \in (-1, 1)$ and assume, for contradiction, that

$$f_{\mathrm{c}}\big(\boldsymbol{x}_0 - \tfrac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'\big) = f_{\mathrm{c}}\big(\boldsymbol{x}_0 + r\boldsymbol{e} + t\boldsymbol{e}'\big) \quad \text{for all } t \in [-\tfrac{1}{4}, \tfrac{1}{4}]. \tag{34}$$

Both $\|\boldsymbol{x}_1(t) - \boldsymbol{x}_0\|_2^2 = \frac{1}{16} - \frac{t}{2}\mu + t^2 \ge \frac{1}{16}(1-\mu^2)$ and $\|\boldsymbol{x}_2(t) - \boldsymbol{x}_0\|_2^2 = r^2 + 2rt\mu + t^2 \ge r^2(1-\mu^2)$, are bounded below by positive constants on $\mathbb{R}$, so both norms are real-analytic functions of t on $\mathbb{R}$; the κ-terms are polynomials in t. By the identity theorem for real-analytic functions, (34) holds for every $t \in \mathbb{R}$.

Write $\boldsymbol{e} = (e_1, e_2)$ and $\boldsymbol{e}' = (e_1', e_2')$. Expanding the cubes (the t^3-terms cancel),

$$\begin{aligned}\kappa\big(-\tfrac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'\big) - \kappa\big(r\boldsymbol{e} + t\boldsymbol{e}'\big) = &-3\big(\tfrac{1}{4} + r\big)\big((e_1')^2 e_1 + 2(e_2')^2 e_2\big)t^2 \\ &+ 3\big(\tfrac{1}{16} - r^2\big)\big(e_1' e_1^2 + 2e_2' e_2^2\big)t - \big(\tfrac{1}{64} + r^3\big)(e_1^3 + 2e_2^3).\end{aligned} \tag{35}$$

a polynomial in t of degree at most two. By (34) and the definition (27) of f_{c},

$$\kappa\big(-\tfrac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'\big) - \kappa\big(r\boldsymbol{e} + t\boldsymbol{e}'\big) = -10\Big[\big\|-\tfrac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'\big\|_2 - \big\|r\boldsymbol{e} + t\boldsymbol{e}'\big\|_2\Big] \quad \text{for all } t \in \mathbb{R}. \tag{36}$$

By the reverse triangle inequality, $D(t) := \|-\frac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'\|_2 - \|r\boldsymbol{e} + t\boldsymbol{e}'\|_2 \le \big\|-\frac{1}{4}\boldsymbol{e} - r\boldsymbol{e}\big\|_2 = \frac{1}{4} + r$ is bounded on $\mathbb{R}$. A polynomial that is bounded on $\mathbb{R}$ is constant, so the quadratic coefficient in (35) vanishes; since $\frac{1}{4} + r > 0$,

$$(e_1')^2 e_1 + 2(e_2')^2 e_2 = 0. \tag{37}$$

Moreover, (36) now shows that $D(t)$ is constant in t. Rationalizing,

$$D(t) = \frac{\frac{1}{16} - r^2 - 2t\big(\frac{1}{4} + r\big)\mu}{\|-\frac{1}{4}\boldsymbol{e} + t\boldsymbol{e}'\|_2 + \|r\boldsymbol{e} + t\boldsymbol{e}'\|_2},$$

whose limits as $t \to +\infty$ and $t \to -\infty$ are $-(\frac{1}{4}+r)\mu$ and $(\frac{1}{4}+r)\mu$, respectively. Since D is constant, these limits agree, so $\mu = 0$: $\boldsymbol{e} \perp \boldsymbol{e}'$. Then $D(t) = \sqrt{\frac{1}{16}+t^2} - \sqrt{r^2+t^2}$ tends to 0 as $|t| \to \infty$, hence $D \equiv 0$, and $t = 0$ gives $r = \frac{1}{4}$. The right-hand side of (36) is now identically zero, so the polynomial (35) vanishes identically; its constant term with $r = \frac{1}{4}$ gives $-\frac{1}{32}(e_1^3 + 2e_2^3) = 0$, i.e.

$$e_1^3 + 2e_2^3 = 0. \tag{38}$$

Since $\boldsymbol{e} \perp \boldsymbol{e}'$ are unit vectors, $\boldsymbol{e}' = \pm(-e_2, e_1)$, and (37) becomes $e_2^2 e_1 + 2e_1^2 e_2 = e_1 e_2 (e_2 + 2e_1) = 0$. If $e_1 = 0$ then $e_2 = \pm 1$ and $e_1^3 + 2e_2^3 = \pm 2$; if $e_2 = 0$ then $e_1 = \pm 1$ and $e_1^3 + 2e_2^3 = \pm 1$; if $e_2 = -2e_1$ then $e_1 \neq 0$ and $e_1^3 + 2e_2^3 = -15 e_1^3 \neq 0$. All three cases contradict (38). Hence (34) is impossible, which proves (33).

Conclusion. Since $\mathcal{K}$ is compact and $\mathcal{M}$ is continuous, define $s := \min_{\mathcal{K}} \mathcal{M} > 0$. Hence every $\Phi \in \mathcal{N}((2,1);\phi)$ satisfies $\|f_{\mathrm{c}} - \Phi\|_{L^\infty([0,1]^d)} \geq \varepsilon_*$, where $\varepsilon_* := \min\left\{\frac{1-\theta}{8}, \frac{1}{50}, \frac{s}{2}\right\} > 0$.

Step 5: rescale f_{c} into the Hölder class. The map $\boldsymbol{y} \mapsto \|\boldsymbol{y}\|_2$ is $\sqrt{2}$-Lipschitz with respect to $\|\cdot\|_\infty$. Moreover, $\|\nabla \kappa(\boldsymbol{y})\|_1 = 3y_1^2 + 6y_2^2 \leq \frac{9}{4}$ for $\boldsymbol{y} \in \left[-\frac{1}{2}, \frac{1}{2}\right]^2$. Thus f_{c} is Lipschitz with respect to $\|\cdot\|_\infty$, with constant $\sqrt{2} + \frac{9}{40} < 2$. Also, (28) gives $\|f_{\mathrm{c}}\|_{L^\infty([0,1]^2)} \leq \frac{1+\theta}{\sqrt{2}} < 1$. Set $\gamma' := \min\left\{\frac{\lambda}{2}, R\right\}$, $\widetilde{f}_{\mathrm{c}} := \gamma' f_{\mathrm{c}}$. Then $\|\widetilde{f}_{\mathrm{c}}\|_{L^\infty([0,1]^2)} \leq R$. Furthermore, since $\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq 1$ on $[0,1]^2$, $|\widetilde{f}_{\mathrm{c}}(\boldsymbol{x}) - \widetilde{f}_{\mathrm{c}}(\boldsymbol{x}')| \leq 2\gamma' \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \lambda \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \lambda \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty^\alpha$. Therefore $\widetilde{f}_{\mathrm{c}} \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^2)$. Finally, $\mathcal{N}((2,1);\phi)$ is invariant under multiplication of the output layer by the positive constant γ'. Hence

$$\inf_{\Phi \in \mathcal{N}((2,1);\phi)} \|\widetilde{f}_{\mathrm{c}} - \Phi\|_{L^\infty([0,1]^2)} = \gamma' \inf_{\Phi \in \mathcal{N}((2,1);\phi)} \|f_{\mathrm{c}} - \Phi\|_{L^\infty([0,1]^2)} \geq \gamma' \varepsilon_* := \varepsilon_2 > 0.$$

□

E.2 L^p APPROXIMATION WITH $d+1$ HIDDEN NEURONS

In this appendix, we construct a continuous activation with which a network of hidden-layer widths $(d, 1)$ approximates every $f \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$ to accuracy ε in $L^p([0,1]^d)$. The idea of the continuous activation is to smooth the jumps of ϱ over transition intervals whose lengths decay geometrically along the negative axis, and to shift the arguments of the first hidden layer so far into the negative axis that the transition intervals there have negligible total length.

Continuous activation. The activation ϱ in (10) is constant on each interval $[k, k+1)$, $k \in \mathbb{Z}$, and hence can only jump at integers. For $k \in \mathbb{Z}$, let $w_k := 2^{\min\{k,0\}-1} \in (0, \frac{1}{2}]$, i.e. $w_k = \frac{1}{2}$ for $k \geq 0$ and $w_k = 2^{k-1}$ for $k < 0$, and define

$$\varrho_c(t) := \begin{cases} \varrho(t), & t \notin \bigcup_{k\in\mathbb{Z}} (k - w_k,\, k), \\ \varrho(k-1) + \big(\varrho(k) - \varrho(k-1)\big) \dfrac{t - (k - w_k)}{w_k}, & t \in (k - w_k,\, k),\ k \in \mathbb{Z}; \end{cases}$$

see Figure 5. Here $\varrho_c \in C(\mathbb{R})$, $\varrho_c(k) = \varrho(k)$ for every $k \in \mathbb{Z}$, and on each $[k-1, k]$ the values of ϱ_c lie between $\varrho(k-1)$ and $\varrho(k)$. In particular, ϱ_c is locally bounded, $\lfloor t \rfloor \leq \varrho_c(t) \leq \lfloor t \rfloor + 1$ for $t < 0$, and $0 \leq \varrho_c(t) \leq n$ on $[S_n, S_{n+1}]$. Only the negative axis is reached by the first hidden layer below, and there the total length of the transition intervals $(k - w_k, k)$ with $k \leq -N$ is $\sum_{k \geq N} 2^{-k-1} = 2^{-N}$ for every $N \in \mathbb{N}_0$. The ramp lengths on the positive axis are immaterial, because the second hidden layer will be evaluated at integers outside a small exceptional set.

Network. Let $p \in [1, \infty)$, $f \in \mathcal{H}^\alpha_{\lambda,R}([0,1]^d)$ and $\varepsilon > 0$. Let $\delta, M, B, K, n, q_\ell, J_f, N_f$ be defined by (5), (6) and (12) with ε replaced by $\varepsilon/2$, so that $\lambda M^{-\alpha} + \delta \leq \varepsilon/2$, and put

$$N_0 := \max\Big\{0,\ \Big\lceil \log_2 d + p \log_2 \frac{2(2R + \delta n)}{\varepsilon} \Big\rceil\Big\} \in \mathbb{N}_0, \tag{39}$$

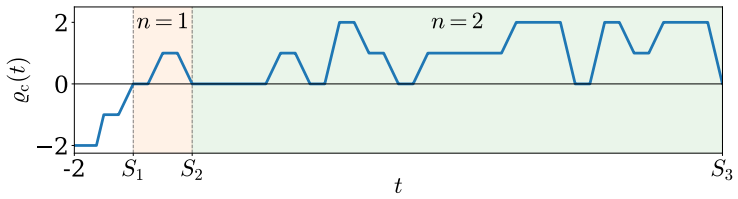


Figure 5: ϱ_c in $[-2, S_3)$, to be compared with ϱ in Figure 2. Each jump of ϱ at an integer k is replaced by a linear ramp on $(k - w_k, k)$, of length $\frac{1}{2}$ for $k \geq 0$ and of geometrically shrinking length 2^{k-1} for $k < 0$.

so that $d\,2^{-N_0}(2R + \delta n)^p \leq (\varepsilon/2)^p$. The network $\Phi^{\varrho_c}_{f,\varepsilon,p} \in \mathcal{N}((d, 1); \varrho_c)$ is defined by

$$u_j(\boldsymbol{x}) := \varrho_c\big(Mx_j - (M+1) - N_0\big), \quad j = 1, \ldots, d, \qquad \text{(1st hidden layer)} \quad (40)$$

$$z(\boldsymbol{x}) := \varrho_c\Big(N_f + \sum_{j=1}^{d} (M+1)^{j-1}\big(u_j(\boldsymbol{x}) + M + 1 + N_0\big)\Big), \qquad \text{(2nd hidden layer)} \quad (41)$$

$$\Phi^{\varrho_c}_{f,\varepsilon,p}(\boldsymbol{x}) := -R + \delta\, z(\boldsymbol{x}). \qquad \text{(output layer)} \quad (42)$$

Compared with (13)–(15), the only changes are the activation and the shift N_0, which is compensated in the second-layer bias; now the second-layer bias is the integer $N_f + (M + 1 + N_0)\sum_{j=1}^{d}(M + 1)^{j-1} = N_f + (M + 1 + N_0)(K - 1)/M$.

Transition region. For $\boldsymbol{x} \in [0, 1]^d$, the arguments $t_j(\boldsymbol{x}) := Mx_j - (M + 1) - N_0$ in (40) lie in $[-(M + 1) - N_0, -1 - N_0]$, so the transition intervals that can be met are $(k - w_k, k)$ with $k \in \{-M - N_0, \ldots, -1 - N_0\}$, each of length $w_k \leq 2^{-N_0-2}$ in t_j, i.e. w_k/M in x_j. Define $\Lambda := \big\{\boldsymbol{x} \in [0, 1]^d : \; t_j(\boldsymbol{x}) \in \bigcup_{k\in\mathbb{Z}}(k - w_k, k) \text{ for some } j\big\}$, whose Lebesgue measure satisfies

$$|\Lambda| \leq d \sum_{k \leq -1-N_0} \frac{w_k}{M} = \frac{d\,2^{-N_0-1}}{M} \leq d\,2^{-N_0}. \quad (43)$$

Lemma E.2. *For $\boldsymbol{x} \in [0, 1]^d \setminus \Lambda$, $u_j(\boldsymbol{x}) = \lfloor Mx_j \rfloor - (M + 1) - N_0$ for all j and $z(\boldsymbol{x}) = q_{r(\boldsymbol{x})}$; hence $\Phi^{\varrho_c}_{f,\varepsilon,p}(\boldsymbol{x}) = \widehat{f}_{r(\boldsymbol{x})}$ and $|f(\boldsymbol{x}) - \Phi^{\varrho_c}_{f,\varepsilon,p}(\boldsymbol{x})| < \varepsilon/2$. For all $\boldsymbol{x} \in [0, 1]^d$, $0 \leq z(\boldsymbol{x}) \leq n$ and $|f(\boldsymbol{x}) - \Phi^{\varrho_c}_{f,\varepsilon,p}(\boldsymbol{x})| \leq 2R + \delta n$.*

Proof. If $\boldsymbol{x} \notin \Lambda$, each $t_j(\boldsymbol{x})$ is negative and outside every transition interval, so $\varrho_c(t_j(\boldsymbol{x})) = \varrho(t_j(\boldsymbol{x})) = \lfloor t_j(\boldsymbol{x}) \rfloor = \lfloor Mx_j \rfloor - (M + 1) - N_0$. The argument of ϱ_c in (41) is then the integer $N_f + r(\boldsymbol{x})$, at which ϱ_c agrees with ϱ, and $\varrho(N_f + r(\boldsymbol{x})) = q_{r(\boldsymbol{x})}$ by (11); the error bound is (7) with $\varepsilon/2$.

For arbitrary $\boldsymbol{x}$, since $t_j(\boldsymbol{x}) \in [-(M + 1) - N_0, -1 - N_0]$ and ϱ_c lies between the neighboring values of $\lfloor\cdot\rfloor$ on this range, $u_j(\boldsymbol{x}) + M + 1 + N_0 \in [0, M]$. Hence the preactivation part in (41) lies in $[N_f,\, N_f + M\sum_{j=1}^{d}(M + 1)^{j-1}] = [N_f,\, N_f + K - 1] \subset [S_n, S_{n+1} - 1]$, because $K \leq n$ and $N_f + n - 1 \leq S_n + n(n + 1)^n - 1 = S_{n+1} - 1$. On $[S_n, S_{n+1}]$ we have $0 \leq \varrho_c \leq n$, so $0 \leq z(\boldsymbol{x}) \leq n$, $\Phi^{\varrho_c}_{f,\varepsilon,p}(\boldsymbol{x}) \in [-R, -R + \delta n]$, and since $f(\boldsymbol{x}) \in [-R, R]$ the last claim follows. □

Theorem E.3 (Fixed continuous activation, $(d, 1)$ architecture, L^p guarantee)**.** *Let $d \in \mathbb{N}_+$, $\alpha \in (0, 1]$, $\lambda > 0$, $R \in \mathbb{N}_+$ and $p \in [1, \infty)$. For every $f \in \mathcal{H}^{\alpha}_{\lambda,R}([0, 1]^d)$ and every $\varepsilon > 0$, the network $\Phi^{\varrho_c}_{f,\varepsilon,p} \in \mathcal{N}((d, 1); \varrho_c)$ defined by* (40)–(42) *satisfies*

$$\big\|f - \Phi^{\varrho_c}_{f,\varepsilon,p}\big\|_{L^p([0,1]^d)} \leq \varepsilon.$$

The activation ϱ_c is continuous and does not depend on f, ε, p or d.

Proof. Splitting $[0,1]^d$ into Λ and its complement and using (43) and Lemma E.2,

$$\begin{aligned}\big\|f-\Phi^{\varrho_c}_{f,\varepsilon,p}\big\|^p_{L^p([0,1]^d)} &= \int_{[0,1]^d\setminus\Lambda}\big|f-\Phi^{\varrho_c}_{f,\varepsilon,p}\big|^p+\int_{\Lambda}\big|f-\Phi^{\varrho_c}_{f,\varepsilon,p}\big|^p\\ &\le \Big(\frac{\varepsilon}{2}\Big)^p + d\,2^{-N_0}(2R+\delta n)^p \le 2\Big(\frac{\varepsilon}{2}\Big)^p \le \varepsilon^p.\end{aligned}$$

□

Remark E.4 (Parameters and bit complexity of $\Phi^{\varrho_c}_{f,\varepsilon,p}$). For $R\in\mathbb{N}_+$, all parameters of $\Phi^{\varrho_c}_{f,\varepsilon,p}$ are integers except the dyadic output weight δ, and ϱ_c takes rational values at rational arguments, so the network can be evaluated in exact arithmetic as in Section 6. Compared with $\Phi^{\varrho}_{f,\varepsilon/2}$, only two parameters change: the first-layer biases decrease by N_0 and the second-layer bias increases by the integer $N_0(K-1)/M$. Since $\delta n\le \delta K+\delta(B-1)$ with $\delta K=\Theta(\varepsilon^{1-d/\alpha})$ and $\delta(B-1)\le 4R+2\delta$ by (5), (39) gives $N_0=O\big(p\,\frac{d}{\alpha}\log\frac{1}{\varepsilon}\big)$, so these changes cost $O(\log\frac{1}{\varepsilon})$ bits for fixed d,α,λ,R,p, while the dominant term $\mathrm{bit}(N_f)=\Theta(\varepsilon^{-d/\alpha}\log\frac{1}{\varepsilon})$ from the proof of Theorem 5.1 is unchanged. Hence, for $d\ge 2$, $\sup_{f\in\mathcal{H}^\alpha_{\lambda,R}([0,1]^d)}\mathrm{bit}\big(\Phi^{\varrho_c}_{f,\varepsilon,p}\big)=\Theta\big(\varepsilon^{-d/\alpha}\log\frac{1}{\varepsilon}\big)$: the continuous L^p construction has the same asymptotic bit complexity as the discontinuous L^∞ construction.

F Additional Figures and Tables

In this appendix, we present additional figures and tables that help to understand our construction or numerical results.

Table 2: Position encoding example for $d=2$ and $M=3$, where $I_1=[0,\frac{1}{3})$, $I_2=[\frac{1}{3},\frac{2}{3})$, $I_3=[\frac{2}{3},1)$, and $I_4=\{1\}$.

Q_ℓ	$\boldsymbol{x}^{(\ell)}$	ℓ	Q_ℓ	$\boldsymbol{x}^{(\ell)}$	ℓ	Q_ℓ	$\boldsymbol{x}^{(\ell)}$	ℓ	Q_ℓ	$\boldsymbol{x}^{(\ell)}$	ℓ
$I_1\times I_1$	$(0,0)$	$(0)_4$	$I_2\times I_1$	$(\frac{1}{3},0)$	$(1)_4$	$I_3\times I_1$	$(\frac{2}{3},0)$	$(2)_4$	$I_4\times I_1$	$(1,0)$	$(3)_4$
$I_1\times I_2$	$(0,\frac{1}{3})$	$(10)_4$	$I_2\times I_2$	$(\frac{1}{3},\frac{1}{3})$	$(11)_4$	$I_3\times I_2$	$(\frac{2}{3},\frac{1}{3})$	$(12)_4$	$I_4\times I_2$	$(1,\frac{1}{3})$	$(13)_4$
$I_1\times I_3$	$(0,\frac{2}{3})$	$(20)_4$	$I_2\times I_3$	$(\frac{1}{3},\frac{2}{3})$	$(21)_4$	$I_3\times I_3$	$(\frac{2}{3},\frac{2}{3})$	$(22)_4$	$I_4\times I_3$	$(1,\frac{2}{3})$	$(23)_4$
$I_1\times I_4$	$(0,1)$	$(30)_4$	$I_2\times I_4$	$(\frac{1}{3},1)$	$(31)_4$	$I_3\times I_4$	$(\frac{2}{3},1)$	$(32)_4$	$I_4\times I_4$	$(1,1)$	$(33)_4$

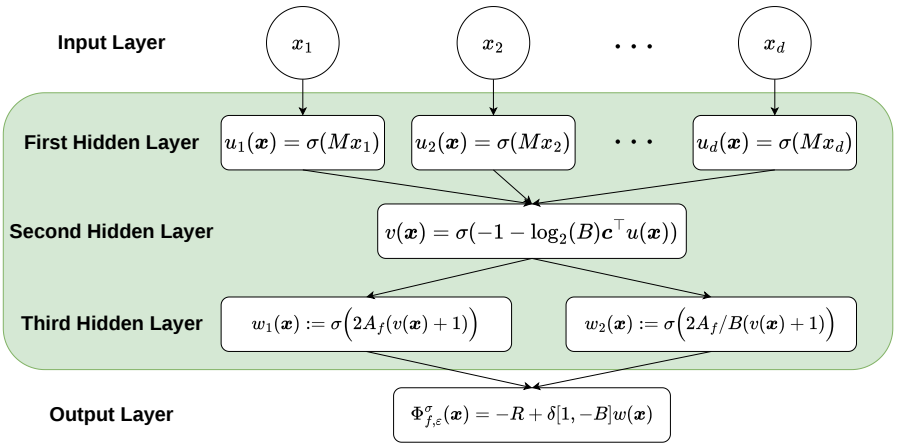


Figure 6: Architecture of the elementary network $\Phi^\sigma_{f,\varepsilon}$ in Theorem 4.1. It has hidden-layer widths $(d,1,2)$ and $d+3$ hidden neurons in total.

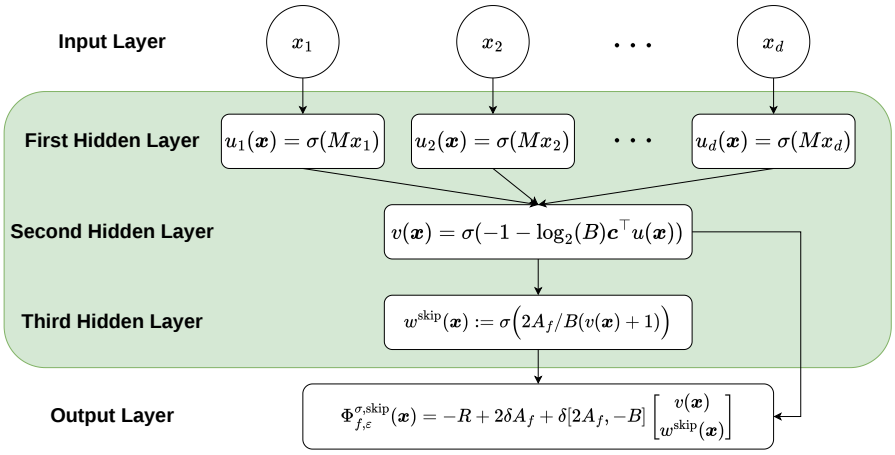


Figure 7: Architecture of the elementary network with a skip connection, $\Phi_{f,\varepsilon}^{\sigma,\text{skip}}$, in Theorem 4.1. It has hidden-layer widths $(d, 1, 1)$ and $d + 2$ hidden neurons in total.

Table 3: Parameters, encoding bit lengths, and empirical errors for the three network constructions on $f_1, \ldots, f_6$. The test set $\mathcal{X}_{\text{test}}$ contains 3,004 points for $f_1, \ldots, f_4$ and 3,008 points for f_5, f_6. The empirical errors are $E_\varrho := \max_{\boldsymbol{x}\in\mathcal{X}_{\text{test}}} |f(\boldsymbol{x}) - \Phi_{f,\varepsilon}^{\varrho}(\boldsymbol{x})|$, $E_\sigma := \max_{\boldsymbol{x}\in\mathcal{X}_{\text{test}}} |f(\boldsymbol{x}) - \Phi_{f,\varepsilon}^{\sigma}(\boldsymbol{x})|$, and $E_{\sigma,\text{skip}} := \max_{\boldsymbol{x}\in\mathcal{X}_{\text{test}}} |f(\boldsymbol{x}) - \Phi_{f,\varepsilon}^{\sigma,\text{skip}}(\boldsymbol{x})|$.

f	ε	δ	M	K	B	$\text{bit}(A_f)$	$\text{bit}(N_f)$	$\frac{\text{bit}(A_f)}{K\log_2 B}$	$\frac{\text{bit}(N_f)}{n\log_2(n+1)}$	E_ϱ	E_σ	$E_{\sigma,\text{skip}}$
f_1	2^{-1}	2^{-2}	13	196	16	782	1 497	0.9974	1.0021	0.3275	0.3275	0.2611
f_1	2^{-2}	2^{-3}	26	729	32	3 643	6 938	0.9995	1.0006	0.1760	0.1760	0.1438
f_1	2^{-3}	2^{-4}	51	2 704	64	16 222	30 834	0.9999	1.0001	0.0912	0.0912	0.0762
f_1	2^{-4}	2^{-5}	101	10 404	128	72 826	138 847	1.0000	1.0000	0.0416	0.0416	0.0335
f_1	2^{-5}	2^{-6}	202	41 209	256	329 670	631 770	1.0000	1.0000	0.0223	0.0223	0.0186
f_1	2^{-6}	2^{-7}	403	163 216	512	1 468 942	2 826 326	1.0000	1.0000	0.0116	0.0116	0.0097
f_1	2^{-7}	2^{-8}	805	649 636	1 024	6 496 358	12 544 008	1.0000	1.0000	0.0056	0.0056	0.0045
f_2	2^{-1}	2^{-2}	49	2 500	32	12 499	28 225	0.9999	1.0002	0.2978	0.2978	0.2785
f_2	2^{-2}	2^{-3}	196	38 809	64	232 853	591 615	1.0000	1.0000	0.1772	0.1772	0.1435
f_2	2^{-3}	2^{-4}	784	616 225	128	4 313 574	11 851 923	1.0000	1.0000	0.0911	0.0911	0.0734
f_3	2^{-1}	2^{-2}	76	5 929	16	23 715	74 316	1.0000	1.0000	0.4500	0.4500	0.4167
f_3	2^{-2}	2^{-3}	152	23 409	32	117 044	339 781	1.0000	1.0000	0.2370	0.2370	0.2169
f_3	2^{-3}	2^{-4}	304	93 025	64	558 149	1 535 414	1.0000	1.0000	0.1189	0.1189	0.1110
f_3	2^{-4}	2^{-5}	608	370 881	128	2 596 166	6 861 527	1.0000	1.0000	0.0584	0.0584	0.0535
f_3	2^{-5}	2^{-6}	1 216	1 481 089	256	11 848 711	30 359 706	1.0000	1.0000	0.0301	0.0301	0.0276
f_4	2^{-1}	2^{-2}	74	5 625	16	22 498	70 078	0.9999	1.0000	0.4219	0.4219	0.3552
f_4	2^{-2}	2^{-3}	147	21 904	32	109 518	315 837	1.0000	1.0000	0.2027	0.2027	0.1654
f_4	2^{-3}	2^{-4}	293	86 436	64	518 614	1 417 500	1.0000	1.0000	0.0993	0.0993	0.0805
f_4	2^{-4}	2^{-5}	586	344 569	128	2 411 981	6 338 158	1.0000	1.0000	0.0465	0.0465	0.0393
f_4	2^{-5}	2^{-6}	1 172	1 375 929	256	11 007 430	28 057 917	1.0000	1.0000	0.0268	0.0268	0.0209
f_5	2^{-1}	2^{-2}	16	4 913	16	19 650	60 249	0.9999	1.0000	0.3384	0.3384	0.2839
f_5	2^{-2}	2^{-3}	32	35 937	32	179 683	543 846	1.0000	1.0000	0.1620	0.1620	0.1399
f_5	2^{-3}	2^{-4}	63	262 144	64	1 572 862	4 718 598	1.0000	1.0000	0.0864	0.0864	0.0755
f_5	2^{-4}	2^{-5}	125	2 000 376	128	14 002 630	41 871 557	1.0000	1.0000	0.0476	0.0476	0.0412
f_6	2^{-1}	2^{-2}	11	1 728	16	6 910	18 588	0.9997	1.0001	0.4025	0.4025	0.3612
f_6	2^{-2}	2^{-3}	22	12 167	32	60 833	165 119	1.0000	1.0000	0.1959	0.1959	0.1717
f_6	2^{-3}	2^{-4}	44	91 125	64	546 748	1 501 341	1.0000	1.0000	0.1092	0.1092	0.0924
f_6	2^{-4}	2^{-5}	88	704 969	128	4 934 781	13 695 580	1.0000	1.0000	0.0517	0.0517	0.0453

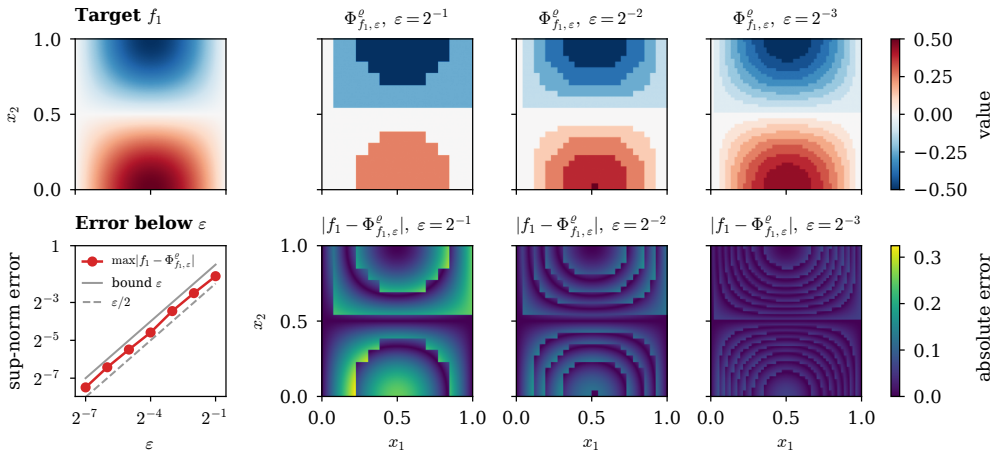


Figure 8: Numerical verification of Theorem 3.1 for the approximation of the 2D function f_1 by the three-neuron network $\Phi^{\varrho}_{f_1,\varepsilon}$. The panels are arranged as in Figure 4, with $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}$ ($M = 13, 26, 51$); the error plot additionally includes finer networks down to $\varepsilon = 2^{-7}$.

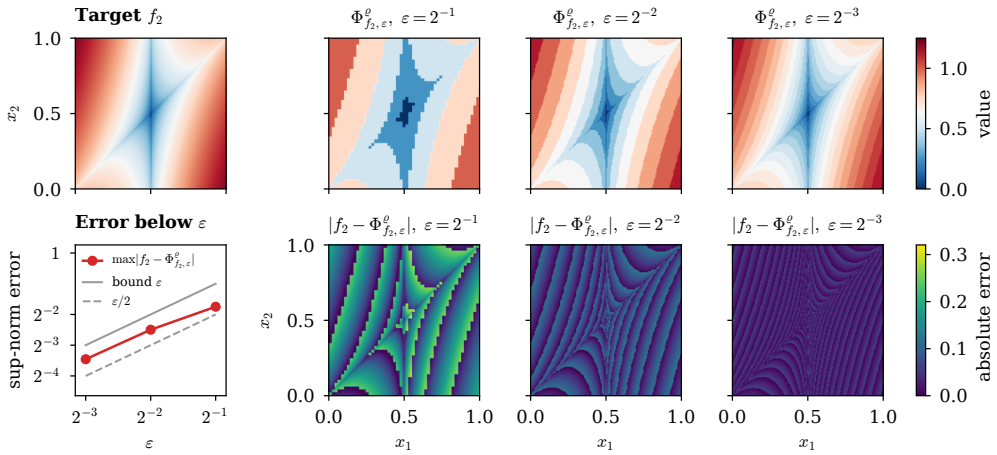


Figure 9: Numerical verification of Theorem 3.1 for the approximation of the 2D function f_2 by the three-neuron network $\Phi^{\varrho}_{f_2,\varepsilon}$. The panels are arranged as in Figure 4, with $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}$ ($M = 49, 196, 784$).

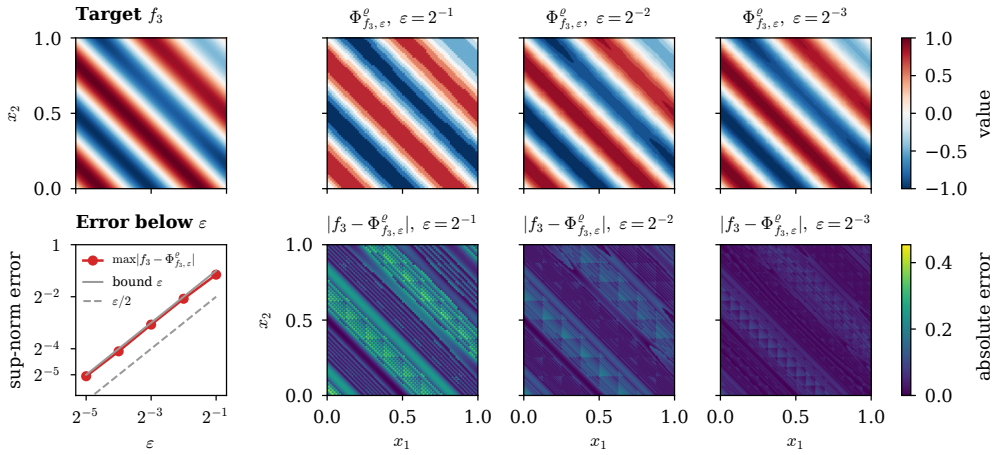


Figure 10: Numerical verification of Theorem 3.1 for the approximation of the 2D function f_3 by the three-neuron network $\Phi^{\varrho}_{f_3,\varepsilon}$. The panels are arranged as in Figure 4, with $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}$ ($M = 76, 152, 304$); the error plot additionally includes finer accuracy down to $\varepsilon = 2^{-5}$.

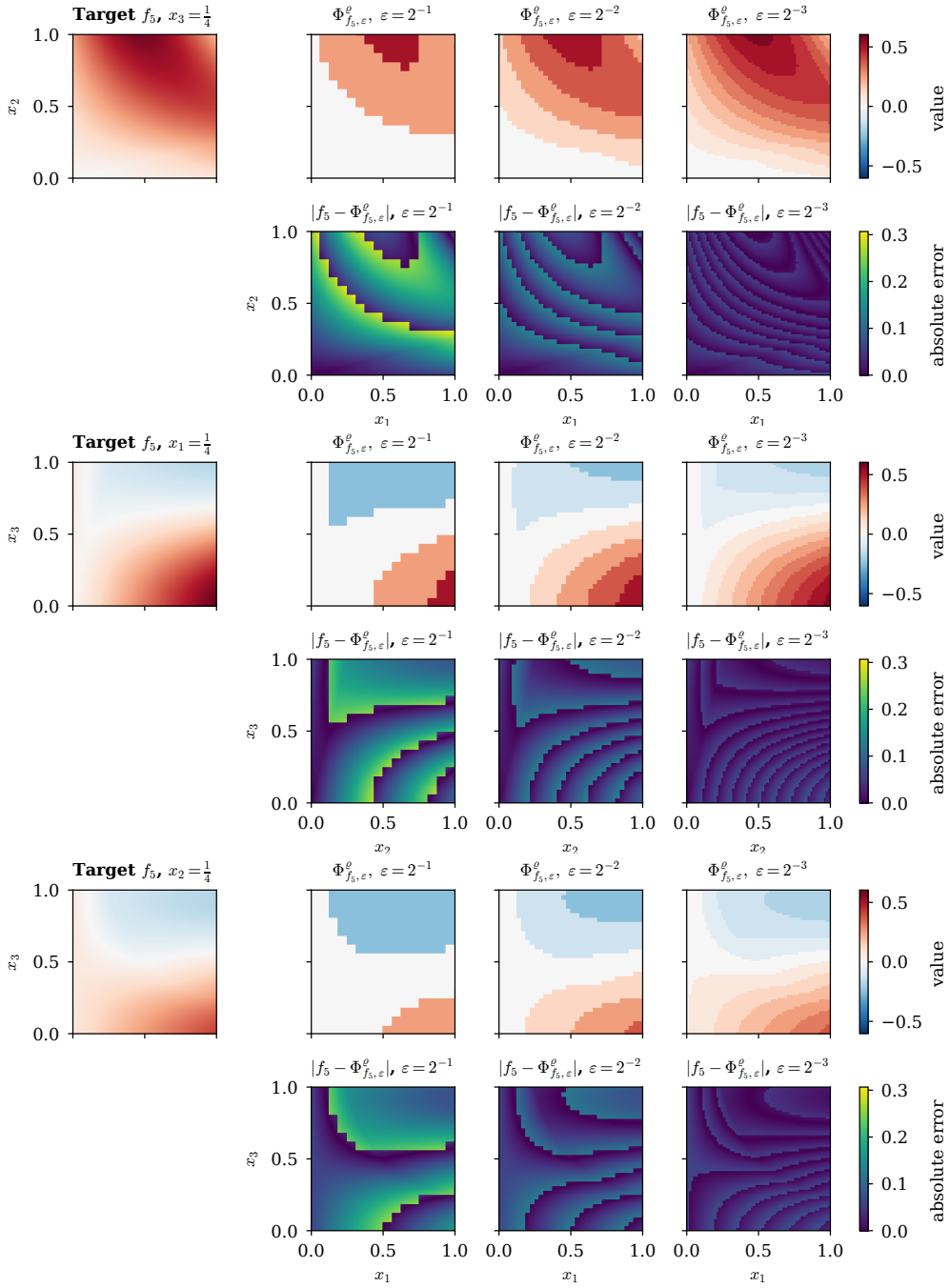


Figure 11: Numerical verification of Theorem 3.1 for the approximation of the 3D function f_5 by the four-neuron network $\Phi^{\varrho}_{f_5,\varepsilon}$. From top to bottom, the panels show the x_1x_2, x_2x_3, and x_1x_3 sections, with the remaining coordinate fixed at $1/4$. Within each panel, the top row contains the target and the network outputs for $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}$ ($M = 16, 32, 63$), and the bottom row contains the corresponding absolute errors.

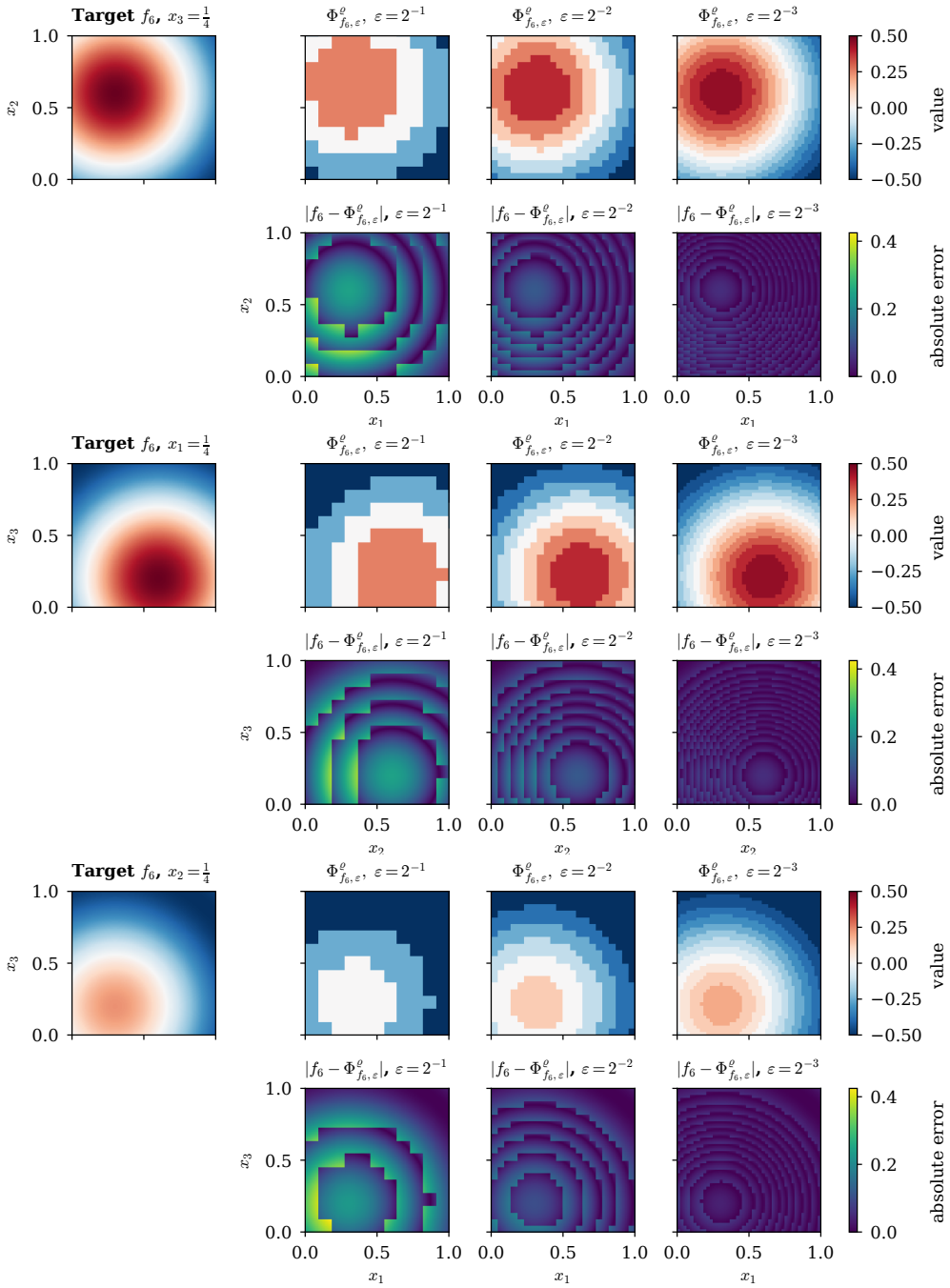
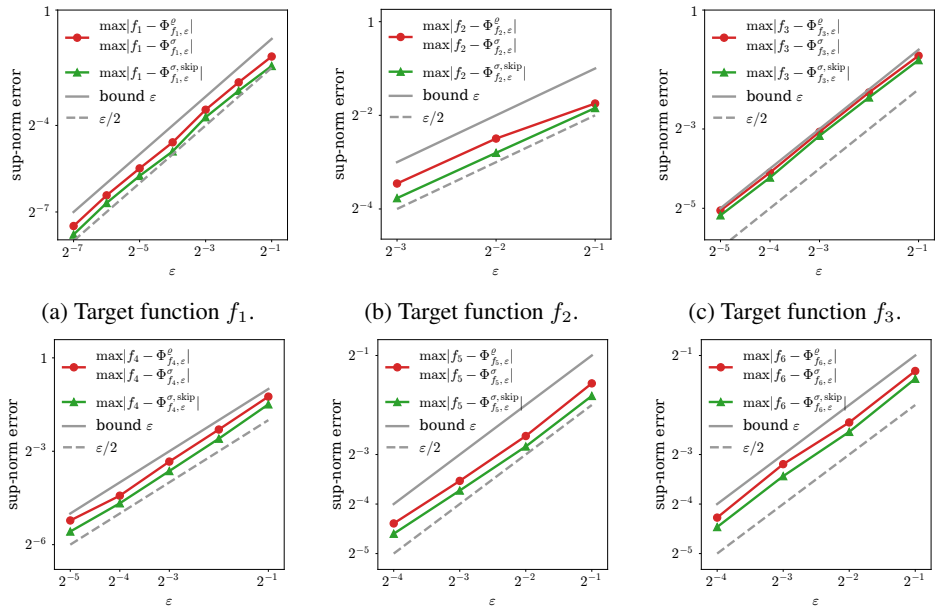


Figure 12: Numerical verification of Theorem 3.1 for the approximation of the 3D function f_6 by the four-neuron network $\Phi^{\varrho}_{f_6,\varepsilon}$. The sections and panels are arranged as in Figure 11, with $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}$ and $M = 11, 22, 44$.



(a) Target function f_1. (b) Target function f_2. (c) Target function f_3.

(d) Target function f_4. (e) Target function f_5. (f) Target function f_6.

Figure 13: Sup-norm errors for the three network constructions applied to $f_1, \ldots, f_6$, evaluated in exact arithmetic over 3,004 test points for $f_1, \ldots, f_4$ ($d = 2$) and 3,008 test points for f_5 and f_6 ($d = 3$). Each test set consists of 2,000 random points, 1,000 boundary points, and the 2^d corners. The exact and elementary networks attain identical maximal errors and are therefore represented by a single curve, while the skip-connection network is shown separately. All errors remain below the prescribed accuracy ε.